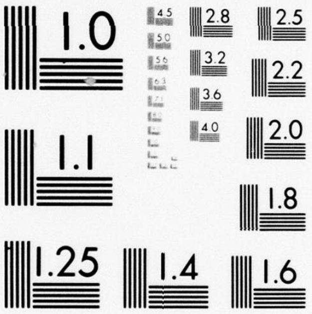
AD-A064 433   ARMY RESEARCH OFFICE RESEARCH TRIANGLE PARK N C   F/G 12/1
TRANSACTIONS OF THE CONFERENCE OF ARMY MATHEMATICIANS(24TH) HEL--ETC(U)
JAN 79

UNCLASSIFIED   ARO-79-1   NL

1 OF 6
AD
A064433



MICROCOPY RESOLUTION TEST CHART

NATIONAL BUREAU OF

ARO Report 79-1

(14)

DwB

TRANSACTIONS OF THE TWENTY-FOURTH CONFERENCE OF ARMY MATHEMATICIANS

Approved for public release; distribution unlimited. The findings in this report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

Sponsored by

The Army Mathematics Steering Committee

on behalf of

THE CHIEF OF RESEARCH, DEVELOPMENT

AND ACQUISITION

(9) Interim technical rept.

(14)

U. S. ARMY RESEARCH OFFICE

(14) ARO-

Report No. 79-1

(11) January 1979

(12) 517p.

(6)

TRANSACTIONS OF THE ~~TWENTY-FOURTH~~ CONFERENCE OF ARMY MATHEMATICIANS (24th)

held 31 May and 1-2 June 1978, University of Virginia, Charlottesville.

Sponsored by the Army Mathematics Steering Committee

Hosts

U. S. Army Foreign Science and Technology Center
with the
School of Engineering and Applied Science
University of Virginia
Charlottesville, Virginia

31 May and 1-2 June 1978

1473

Approved for public release; distribution unlimited. The findings in this report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

U. S. Army Research Office
P. O. Box 12211
Research Triangle Park, North Carolina

FOREWORD

The theme of the Twenty-fourth Conference of Army Mathematicians was "stochastic processes". Four of the six invited speakers, listed below, spoke on topics related to this theme. In recent years there has been a shift of interest from the deterministic descriptive processes to the stochastic processes. This has been brought about by the need to explain many of the phenomena arising in such fields as physics, engineering, biology, and medicine. The complexities and uncertainties that appear in these fields have forced mathematicians to make frequent use of probabilistic concepts. Army scientists are having to deal with stochastic equations, principally those associated with ordinary and partial differential equations. These are concerned with such phenomena as wave propagation, turbulence and diffusion theory.

Speaker and Institution	Area of Talk
Professor E. J. McShane University of Virginia	Choosing a Mathematical Model for a System Affected by Noise
Professor R. E. Kalman University of Florida	Nonlinear Realization Theory
Professor Y. K. Lin University of Illinois	Stochastic Theory of Rotor Blade Dynamics
Professor Roger Brockett Harvard University	Optimal Multilinear Estimators
Professor Ronald DiPerna Mathematics Research Center University of Wisconsin-Madison	Hyperbolic Conservation Laws
Professor Eugene Wong University of California-Berkeley	A Martingale Theory of Random Fields

The Twenty-fourth Conference of Army Mathematicians was held 31 May - 2 June 1978 at Charlottesville, Virginia. The U. S. Army Foreign Science and Technology Center (AFSATC), together with the School of Engineering and Applied Sciences of the University of Virginia, served as its hosts. Colonel Anthony P. Simkus, Commanding Officer of the US Army Research Office, played a key role in obtaining the hosts for this meeting. This fact is borne out by the following quotation from a letter by Colonel Claire J. Reeder, Commanding Officer of AFSATC. "I was pleased to receive the proposal by your office to hold the

24th Conference of Army Mathematicians in Charlottesville. As another Army Organization with a scientific and technical mission, I welcome such opportunities to interact with the Army research community. In this case the University of Virginia will be cooperating with us as joint host for the meeting and will provide the conference facilities."

This conference is part of a continuing program of Army-wide symposia held under the auspices of the Army Mathematics Steering Committee (AMSC) to promote better communication among Army scientists. In order that this mission be accomplished, a large number of individuals must expend a great deal of effort. It is not possible to single out all the persons involved in making the 1978 conference such a scientific success, but members of the AMSC would like to recognize a few of these individuals as well as certain organizations. First of all they would like to express their gratitude to the University of Virginia and the AFSATC for providing the necessary facilities and the cordial atmosphere for this conference. Special recognition is due the outstanding arrangements made possible by the two chairpersons on Local Arrangements. Mrs. Betty Jane Pruffer who handled, without a hitch, the administrative details and Mr. Kent Schlussel who handled in a similar matter the technical problems. Finally, the members of the AMSC would like to commend both the invited speakers and the authors of contributed papers for their excellent presentations and the valuable contributions of their papers to the field of science.

TABLE OF CONTENTS*

*This table of contents contains only the papers that are published in this technical manual. For a list of all papers presented at the Twenty-fourth Conference of Army Mathematicians, see the Program of the meeting.

TITLE PAGE

TITLE PAGE

PROGRAM

WEDNESDAY
31 May 1978

0800-0830 Bus from Holiday Inn to Mechanical Engineering Building

0815-0845 Registration

0845-0900 Opening Remarks

0900-1000 GENERAL SESSION I

CHAIRPERSON - A. S. Galbraith, Durham, North Carolina

SPEAKER - E. J. McShane, Charlottesville, Virginia
TITLE - CHOOSING A MATHEMATICAL MODEL FOR A SYSTEM AFFECTED BY NOISE

1000-1030 Break

1030-1130 TECHNICAL SESSION I

CHAIRPERSON - C. C. White, University of Virginia

SOME BOUNDS FOR OPTIMAL MANEUVERS AND PREDICTORS

Harry L. Reed, Jr., Ballistic Research Laboratory, Aberdeen Proving Ground, Maryland

STOCHASTIC MODELS AND TIME-VARIABLE DETERMINISTIC MODELS OF COMBAT

Roger F. Willis, US Army TRASANA, White Sands Missile Range, New Mexico

ON SOLUTIONS OF NONCOOPERATIVE GAMES: AN AXIOMATIC APPROACH

Prakash P. Shenoy, Mathematics Research Center, University of Wisconsin-Madison

1030-1130 TECHNICAL SESSION II

CHAIRPERSON - Earl C. Steeves, US Army Natick R&D Command, Natick, Massachusetts

IMPROVEMENTS IN THE COMPUTED INTERNAL ENERGIES FOR SHAPED CHARGES

James A. Schmitt, Ballistic Research Laboratory, Aberdeen Proving Ground, Maryland

STRESS INTENSITY FACTORS FOR A CIRCULAR RING WITH UNIFORM ARRAY OF RADIAL CRACKS USING CUBIC ISOPARAMETRIC SINGULAR ELEMENTS

S. L. Pu and M. A. Hussain, Watervliet Arsenal, Watervliet, New York

TEMPERATURES AND STRESSES DUE TO QUENCHING OF HOLLOW CYLINDERS

John D. Vasilakis, Watervliet Arsenal, Watervliet, New York

1130-1300 Lunch

1300-1500 <u>TECHNICAL SESSION III</u>

CHAIRPERSON - N. P. Coleman, US Army Armament R&D Command

PARITY AND SYMMETRY IN LINEAR DIFFERENTIAL SYSTEMS

Leon Kotin, US Army Communications R&D Command, Fort Monmouth, New Jersey

CONSERVATION SOLUTIONS OF ONE-DIMENSIONAL NONLINEAR DISTRIBUTION WITH DRIFT

Siegfried H. Lehnigk, MIRADCOM, Redstone Arsenal, Alabama

ASYMPTOTIC SOLUTIONS TO A STABILITY PROBLEM

David A. Peters, Washington University, and Julian J. Wu, Watervliet Arsenal, Watervliet, New York

INVERSE BOUNDARY VALUE PROBLEMS

William Symes, Mathematics Research Center, University of Wisconsin-Madison

SOME ANALYTICAL ASPECTS OF A NONLINEAR TRANSIENT ELECTROMAGNETIC FIELD PENETRATION PROBLEM IN A SEMI-INFINITE MEDIUM

W. J. Croisant and P. Nielson, Construction Engineering Research Laboratory

FORMULATION OF BOUNDARY INTEGRAL EQUATION METHODS VIA GREEN'S THEOREM

Ben Noble, Mathematics Research Center, University of Wisconsin-Madison

1300-1500 TECHNICAL SESSION IV

CHAIRPERSON - San Li Pu, Watervliet Arsenal, Watervliet,
New York

IN-PLANE DEFORMATION OF THE NON-COAXIAL PLASTIC SOIL

Shunsuke Takagi, US Army Cold Regions Research and Engineering Laboratory, Hanover, New Hampshire

LARGE PLASTIC DEFORMATION IN A RADIAL DRAWING PROCESS

Peter C. T. Chen, Watervliet Arsenal, Watervliet, New York

ON THE LIMITATIONS AND IMPROVEMENT OF PRESENT NUMERICAL WEATHER PREDICTION

H. Baussus von Luetzow, US Army Engineer Topographic Laboratories, Fort Belvoir, Virginia

MAJORIZATION FORMULAS FOR A BIHARMONIC FUNCTION OF TWO VARIABLES

J. Barkley Rosser, Mathematics Research Center, University of Wisconsin-Madison

A NUMERICAL METHOD FOR LARGE STIFF SYSTEMS OF ORDINARY DIFFERENTIAL EQUATIONS

T. P. Coffee, J. M. Heimerl, and M. D. Kregel, Ballistic Research Laboratory, Aberdeen Proving Ground, Maryland

EFFICIENT COMPUTER MANIPULATION OF TENSOR PRODUCTS

Carl de Boor, Mathematics Research Center, University of Wisconsin-Madison

1500-1530 Break

1530-1630 GENERAL SESSION II

CHAIRPERSON - J. B. Rosser, Mathematics Research Center,
University of Wisconsin-Madison

SPEAKER - R. E. Kalman, University of Florida
TITLE - NONLINEAR REALIZATION THEORY

THURSDAY

1 June 1978

0900-1000 TECHNICAL SESSION V

CHAIRPERSON - Edward W. Ross, US Army Natick R&D Command, Natick, Massachusetts

THE RADIATION STRENGTH OF A DIELECTRIC ANTENNA AND ASYMPTOTIC APPROACH

Walter Pressman, US Army Communications R&D Command, Fort Monmouth, New Jersey

THE INTEGRAL EQUATION OF IMAGE RECONSTRUCTION

Louis B. Rall, Mathematics Research Center, University of Wisconsin-Madison

STOCHASTIC VOLTERRA EQUATIONS

Marc E. Berger, Mathematics Research Center, University of Wisconsin-Madison

0900-1000 TECHNICAL SESSION VI

CHAIRPERSON - Larry Gambino, US Army Engineering Topographic Laboratory

ON THE STABILITY OF NONLINEAR INTERPOLATING SPLINES

Michael Golomb, Mathematics Research Center, University of Wisconsin-Madison

ON THE OPTIMIZATION OF THE MODIFIED ENTROPY SPECTRUM OF LINEAR ADAPTIVE FILTERS

Jacob Benson, US Army Communications R&D Command, Fort Monmouth, New Jersey

TIME OPTIMAL REJECTION SEQUENCING

Paul T. Boggs and Robert L. Launer, US Army Research Office, Research Triangle Park, North Carolina

BAND MATRICES WITH TOEPLITZ INVERSES

W. F. Trench and T. N. E. Greville, Mathematics Research Center, University of Wisconsin-Madison

1000-1030 Break

1030-1130 GENERAL SESSION III

CHAIRPERSON - Steven Wolff, Ballistic Research Laboratory, Aberdeen Proving Ground, Maryland

SPEAKER - Y. K. Lin, University of Illinois at Urbana-Champaign
TITLE - STOCHASTIC THEORY OF ROTOR BLADE DYNAMICS

1130-1300 Lunch

1300-1400 GENERAL SESSION IV

CHAIRPERSON - Harry Reed, Ballistic Research Laboratory, Aberdeen Proving Ground, Maryland

SPEAKER - Roger Brockett, Harvard University, Cambridge, Massachusetts
TITLE - OPTIMAL MULTILINEAR ESTIMATORS

1400-1445 Briefing - COL C. Reeder, Commander, FSTC

1445-1515 Break

1515-1615 GENERAL SESSION V

CHAIRPERSON - Leon Kotin, US Army Communications R&D Command, Fort Monmouth, New Jersey

SPEAKER - Ronald DiPerna, Mathematics Research Center, University of Wisconsin-Madison
TITLE - HYPERBOLIC CONSERVATION LAWS

FRIDAY

2 June 1978

0900-1020 TECHNICAL SESSION VII

CHAIRPERSON - James Thompson, US Army TARADCOM,
Warren, Michigan

ANALYSIS OF STOCHASTIC REYNOLDS EQUATION AND RELATED PROBLEMS

P. L. Chow, Wayne State University, Detroit, Michigan
E. A. Saibel, US Army Research Office, Research Triangle Park, North Carolina

A MATHEMATICAL MODEL FOR ENZYMATIC HYDROLYSIS AND FERMENTATION OF CELLULOSE BY TRICHODERMA

Edward W. Ross, Jr., and Nicolai Peitersen, US Army Natick R&D Command, Natick, Massachusetts

A NEW MODEL FOR EVALUATING EFFECTIVENESS OF FRAGMENTING WARHEADS IN DYNAMIC ENCOUNTERS

Edgar A. Cohen, Jr., Naval Surface Weapons Center, Silver Spring, Maryland

BOSE OPERATOR REPRESENTATION FOR GENERAL N-LEVEL SYSTEMS

C. M. Bowden, MIRADCOM, Redstone Arsenal, Alabama
C. A. Coulter, University of Alabama-Tuscaloosa; and
N. M. Witriol, Louisiana Tech University, Ruston

0900-1020 TECHNICAL SESSION VIII

CHAIRPERSON - Ms. Mary Scott, FSTC

GENERATING THE EFFICIENT SET FOR MULTIPLE OBJECTIVE LINEAR PROGRAMS

J. G. Ecker, Rennselaer Polytechnic Institute
Nancy S. Hegner, State University of New York at Albany

APPLICATION OF JENSEN'S INEQUALITY FOR ADAPTIVE SUB-OPTIMAL DESIGN

Chelsea C. White and David P. Harrington, University of Virginia

PROBLEMS WITH SOFTWARE DEVELOPMENT IN THE SOVIET UNION

S. E. Goodman, The Woodrow Wilson School of Public and International Affairs, Princeton University, Princeton, New Jersey

1020-1050 Break

1050-1150 GENERAL SESSION VI

CHAIRPERSON - Walter Pressman, US Army Communications R&D Command

SPEAKER - Eugene Wong, University of California-Berkeley, Berkeley, California
TITLE - A MARTINGALE THEORY OF RANDOM FIELDS

CHOOSING A MATHEMATICAL MODEL FOR A SYSTEM AFFECTED BY NOISE

E. J. McShane
Department of Mathematics
University of Virginia
Charlottesville, Virginia
22903

ABSTRACT. There are several kinds of stochastic integrals, each with its own calculus. No one method of setting up stochastic models is appropriate for making a mathematical idealization of every problem involving random noises. The nature of the system determines the integral, or other limit process, that is suitable for the mathematical model. Three examples are presented. In the first the appropriate stochastic differential equation is in "canonical form"; a Stratonovich equation would serve as well. In the second the appropriate equation involves an Itô integral. The third example is an optimal control problem, in which for each control of a certain special operationally feasible type the response satisfies an equation in canonical form, and the control that can be called optimal for continuously acquired information is a type of weak limit of controls of the special type.

I. INTRODUCTION.

Whenever we use mathematics involving a limit process in an experimental situation we are performing an idealization. A frequently occurring example is the replacement of the sum of a large (but finite) number of terms by an integral whose value is acceptably near to it, the integral being more manageable in theoretical studies and even in computation than a sum of extremely many terms. But every such idealization will lead us into error if pushed too far. A fluid, made of molecules, can be replaced in theoretical discussions by an idealization that is spatially homogeneous, and this is convenient and adequately accurate for hydrodynamics, but cannot be used to draw conclusions about regions of molecular dimensions. However, usually the limits on the use of the idealization are easily grasped, and the idealization works so well that we get into the habit of thinking that it is the physical quantity, and not merely a simplified and not perfectly accurate model of it.

When it became important to study systems affected by random noises, a new type of difficulty arose. These problems often involved at least two types of idealization. A sum was replaced by an integral, and for this to be accurate the time-intervals need to be short; and the random disturbance was replaced by some more manageable stochastic process with nearly the same finite-dimensional distributions, but this approximation is valid only if the time-intervals are long. It is not surprising that deductions that involved both kinds of approximation sometimes produced puzzling results. What I wish to bring out is that at least some of these puzzles disappear if we give the central rôle to those finitely-computable quantities that are within reach of experiment. We shall see that there is no universal answer to such a question as "what kind of stochastic integral should we use in modeling noisy systems ?" . Different kinds of systems require different idealizations. Integrals and differential equations are merely tools needed to make a useful model, and their forms should be dictated by the requirements of the modeling procedure.

II. RAPIDLY RESPONDING SYSTEMS. The first kind of system that we shall consider is the type whose evolution in time, in the absence of noises, is describable with high accuracy by some differential equations. Suppose that the state of the system can be specified by n real numbers $x^1, \cdots, x^n$, and that according to some well-verified theory, in the absence of disturbances these variables satisfy a system of differential equations

$$dx^i/dt = f^i(t,x(t)) \quad (i = 1,\cdots,n). \tag{1}$$

Now let the system be subjected to r disturbances, the accumulated amount of the ρ-th disturbance up to time t being $z^\rho(t)$ $(\rho = 1,\cdots,r)$. Then the intensity of that disturbance at time t is $\dot{z}^\rho(t)$, if the derivative exists. We shall assume that the system is simple enough so that at each time t and state x it responds linearly to the disturbances, the sensitivity of x^i to the ρ-th disturbance being $g^i{}_\rho(t,x)$. Then the system will satisfy the differential equations

$$dx^i/dt = f^i(t,x(t)) + \sum_{\rho=1}^{r} g^i{}_\rho(t,x(t))\, \dot{z}^\rho(t).$$

These can be profitably re-written in the integrated form

$$x^i(t) = x^i(a) + \int_a^t f^i(\tau,x(\tau))\, d\tau + \sum_{\rho=1}^{r} \int_a^t g^i{}_\rho(\tau,x(\tau))\, dz^\rho(\tau), \tag{2}$$

if the z are of bounded variation.

To avoid unessential difficulties we shall assume that the f^i and g^i are three times continuously differentiable. In any specific system there will usually be some physically imposed bound L on the absolute values of the difference quotients of the noises, so that

$$|z^\rho(t) - z^\rho(s)| \leq L(t-s) \quad (a \leq s \leq t \leq b); \tag{3}$$

disturbances that do not satisfy (3) will either be impossible or will wreck the system. So we shall assume that all physically realizable z^ρ satisfy (3) with some constant L. Also, the physically attainable values of the x^i will be bounded. Beyond that bound we can change the f^i and $g^i{}_\rho$ arbitrarily without affecting any realizable solution, so we can and shall assume that the f^i and g^i and all their partial derivatives of first, second and third orders are bounded.

The integrals in (2) are Stieltjes integrals, and can be defined in any of several ways which for Lipschitzian z are all equivalent. We choose this definition. Let δ be positive, and let f and z be real-valued on an interval [a',b'] that contains [a,b]. A "δ-fine partition" $\mathcal{P}$ of [a,b] is defined to be a finite set $(t_0, t_1, \cdots, t_q, \tau_1, \cdots, \tau_q)$ of real numbers such that

$$t_0 = a \leq t_1 \leq \cdots \leq t_q = b \tag{4}$$

and

$$\text{for } j = 1,\cdots,q,\ \tau_j \text{ is in } [a',b'] \text{ and } \tau_j - \delta < t_{j-1} \leq t_j < \tau_j + \delta. \tag{5}$$

The "partition-sum" corresponding to $\mathcal{P}$ has the familiar form

$$S(\mathcal{P};\, f,z) = \sum_{j=1}^{q} f(\tau_j)[z(t_j) - z(t_{j-1})]. \tag{6}$$

If there is a number J such that to each positive ε there corresponds a positive δ for which

$$|S(\mathcal{P};\ f,z) - J| < \varepsilon \tag{7}$$

whenever $\mathcal{P}$ is a δ-fine partition of $[a,b]$, we define

$$\int_a^b f(t)\ dz(t) = J.$$

When z is continuously differentiable this can easily be shown to be equivalent to the ordinary Riemann integral of $f\dot{z}$ from a to b. In particular, when $z(t) = t$ it coincides with the Riemann integral of f.

Corresponding to any one accurately-known disturbance $z = (z^1,\cdots,z^r)$ that satisfies the Lipschitz condition (3) the response of the system is given by (2), with a credibility as good as that of the physical theory underlying the theory of the system. But we seldom are much interested in the response to any one particular disturbance. Usually we have some sort of estimate for the joint distribution of the values of the $z^\rho(t)$ at certain finite subsets

$$t_0 = a < t_1 <\cdots< t_q = b \tag{8}$$

of values of t, and we wish to find out something about the distribution of the end-values $x^i(b)$ produced by these disturbances. The actual noise-functions will be members of some class of functions on $[a,b]$, and there is a probability-measure P defined on a σ-algebra of subsets of that class of functions. But usually the stochastic process thus defined is neither accurately known nor mathematically tractable. We wish to replace it by an idealization in which there is a probability $\tilde{P}$ assigned to the sets that belong to a σ-algebra of subsets of another class of functions on $[a,b]$; the members of this new class we shall denote by $\tilde{z}$. This idealization should be manageable by some definable mathematical procedures, and should produce end-values $\tilde{x}^i(b)$ with nearly the same distribution as that of the $x^i(b)$. For instance, suppose that the increment $z^\rho(t) - z^\rho(s)$ is the sum of small independent disturbances occurring in the interval $[s,t]$, and that these small disturbances have a uniform distribution in time. By the central limit theorem, if the time-interval $[s,t]$ is long enough to include many of these small disturbances the increment $z^\rho(t) - z^\rho(s)$ will have a distribution that is nearly normal, with variance proportional to $t - s$. In other examples too we shall meet the phenomenon that the replacement of the actual process z by an idealization can be made with acceptable accuracy if we use only finite-dimensional distributions corresponding to sufficiently widely spaced times $t_0,\cdots,t_q$, but not if the intervals $[t_{j-1},t_j]$ are short. This implies that such essential information as the value of $x^i(b)$ should be estimated by procedures that provide acceptably close approximations even when the intervals $[t_{j-1},t_j]$ are not very short.

Suppose then that we wish to estimate the solution to (2) using only the values of the $z^\rho(t)$ at times

$$t_0 = a < t_1 <\cdots< t_q = b. \tag{9}$$

The values of $x^i(t)$ on each $[t_{j-1},t_j]$ satisfy

$$x^i(t) - x^i(t_{j-1}) = \int_{t_{j-1}}^t f^i(\tau,x(\tau))d\tau + \sum_{\rho=1}^r \int_{t_{j-1}}^t g^i{}_\rho(\tau,x(\tau))\ dz^\rho(\tau). \tag{10}$$

As a first approximation (Euler's) to x(t) we can replace the integrands in (10) by their values at t_{j-1}, computing the estimates $x_1^i(t_j)$ for the $x^i(t_j)$ by successive applications of the formula

$$x_1^i(t) - x_1^i(t_{j-1}) = f^i(t_{j-1}, x_1(t_{j-1}))(t-t_{j-1}) + \sum_{\rho=1}^{r} g^i_\rho(t_{j-1}, x_1(t_{j-1}))(z^\rho(t) - z^\rho(t_{j-1})). \qquad (11)$$

But this is a poor approximation for $x^i(t_j)$ unless the intervals $[t_{j-1}, t_j]$ are very short. There is a well-known improvement on this procedure (the "modified Euler" method) that for equations (1) yields better results. We shall extend this method to equations (2) to obtain a better approximation x_2, thus. Having computed $x_2(t_{j-1})$, we obtain a first approximation $y(t_j)$ to $x_2(t_j)$ by (11):

$$y^i(t_j) = x_2(t_{j-1}) + f^i(t_{j-1}, x_2(t_{j-1}))\,\Delta_j t + \sum_{\rho=1}^{r} g^i_\rho(t_{j-1}, x_2(t_{j-1}))\,\Delta_j z, \qquad (12)$$

where we have written

$$\Delta_j t = t_j - t_{j-1}, \quad \Delta_j z^\rho = z^\rho(t_j) - z^\rho(t_{j-1}). \qquad (13)$$

Now we form a better estimate for the right member of (10) by replacing the integrands by the average of their values at $(t_{j-1}, x_2(t_{j-1}))$ and at $(t_j, y(t_j))$. This yields

$$x_2^i(t_j) - x_2^i(t_{j-1}) = \frac{1}{2}[f^i(t_{j-1}, x_2(t_{j-1})) + f^i(t_j, y(t_j))]\,\Delta_j t + \frac{1}{2}\sum_{\rho=1}^{r}[g^i_\rho(t_{j-1}, x_2(t_{j-1})) + g^i_\rho(t_j, y(t_j))]\,\Delta_j z^\rho .$$

We can simplify this by applying the theorem of the mean to the terms involving $y(t_j)$ and discarding all terms with three or more factors from the list (13); these tends to 0 with max $(t_j - t_{j-1})$ for all disturbances that we shall consider. The resulting approximation x_3, differing little from x_2, satisfies

$$\begin{aligned} x_3^i(t_j) - x_3^i(t_{j-1}) &= f^i\,\Delta_j t + \frac{1}{2}(\partial f^i/\partial t)(\Delta_j t)^2 \\ &+ \frac{1}{2}\sum_{k=1}^{n}(\partial f^i/\partial x^k)(f^k\,\Delta_j t + \sum_{\rho=1}^{r} g^i_\rho\,\Delta_j z^\rho)\,\Delta_j t \\ &+ \sum_{\rho=1}^{r} g^i_\rho\,\Delta_j z^\rho + \frac{1}{2}\sum_{\rho=1}^{r}(\partial g^i_\rho/\partial t)\,\Delta_j t\,\Delta_j z^\rho \\ &+ \frac{1}{2}\sum_{\rho=1}^{r}\sum_{k=1}^{n}(\partial g^i_\rho/\partial x^k)(f^k\,\Delta_j t + \sum_{\sigma=1}^{r} g^i_\sigma\,\Delta_j z^\sigma)\,\Delta_j z^\rho , \end{aligned} \qquad (14)$$

wherein the f^i, g^i and their partial derivatives are all evaluated at $(t_{j-1}, x_3(t_{j-1}))$. It is well known that if all z^ρ are 0, this is a much better approximation to the solution x of the differential equation (1) than x_1 is. It can also be shown, more tediously, that this remains true when there is just one noise ($r = 1$). When $r > 1$, x_3 is sometimes but not always a great improvement over x_1, depending on the properties of the g^i_ρ . For brevity, we suppress details. But in all such cases, including the important case $r = 1$, we can obtain a required accuracy of approximation to x by using (14) with longer intervals $[t_{j-1}, t_j]$ than the Euler approximation demands.

The second stage of idealization is to interpret the $z^\rho(t_j)$ as the values at the t_j of a function $\tilde{z}^\rho$ of the idealized process, with the probability measure $\tilde{P}$. For each particular set of values $z^\rho(t_j)$ this requires nothing more than replacing each symbol $z^\rho(t_j)$ by $\tilde{z}^\rho(t_j)$, which has the same values; but for computing expectations of functions of the $\tilde{x}^i{}_3(b)$ with the idealized process we must use the probability measure $\tilde{P}$, and the tilde over the z reminds us to do this. For the same reason we shall write $\tilde{x}_3{}^i(b)$ for the estimate of the end-values of x^i as computed by (14), with $\tilde{z}^\rho$ in place of z^ρ.

In a certain sense we are now finished; (14) gives us an approximation $\tilde{x}_3{}^i(b)$ for $x^i(b)$, with a distribution computable by (14) from the probability measure $\tilde{P}$. But this $\tilde{x}_3{}^i(b)$ is a finite sum, and as usual finite sums are harder to work with than integrals. If we add equations (14) for $j = 1,\cdots,q$, in the right member we obtain several types of sums. The first is of the form

$$\sum_{j=1}^{q} \phi(t_{j-1})\,\Delta_j t, \tag{15}$$

where for notational simplicity we have written $\phi(t_{j-1})$ for $f^i(t_{j-1},x_3(t_{j-1}))$. Cauchy defined the integral of ϕ to be the limit of such sums as max (t_j-t_{j-1}) tends to 0. But except for continuous functions this definition proved itself hard to work with. Riemann's definition is technically much superior; it is equivalent to defining $\int\phi\,dt$ as the limit of partition-sums (6) (with ϕ in place of f) for δ-fine partitions P as δ tends to 0. Thus sums (15) will tend to the ordinary Riemann integral of ϕ as δ tends to 0.

Another of the sums is of the form

$$\sum_{j=1}^{q} \phi(t_{j-1})\,\Delta_j\tilde{z}^\rho\ ; \tag{16}$$

and this is not so tractable. Fortunately, we do not need to know that the sum (16) tends to a limit for every function $\tilde{z}^\rho$. It is enough to know that (16) converges in probability (with measure P), and that this happens for all processes in a class large enough to include all those that we have any interest in using, in particular for Wiener processes and for processes that satisfy the Lipschitz condition (3). Such a class can be described as follows:

$(\Omega,\mathcal{A},P)$ is a probability triple, and z a function on $[a,b]\times\Omega$;

$\{\mathcal{F}_s : a \leq s \leq b\}$ is an increasing family of σ-subalgebras of $\mathcal{A}$; (17)

for each t in $[a,b]$, the function $\omega\mapsto z(t,\omega)$ is $\mathcal{F}_s$-measurable;

there exists a real number K such that if $a \leq s \leq t \leq b$, almost surely

$$|E(z(t)-z(s)|\mathcal{F}_s)| \leq K(t-s),$$

$$E([z(t)-z(s)]^2|\mathcal{F}_s) \leq K(t-s),$$

$$E([z(t)-z(s)]^4|\mathcal{F}_s) \leq K(t-s).$$

Now if the random variable $\omega\mapsto\phi(t,\omega)$ is $\mathcal{F}_t$-measurable for each t in $[a,b]$, and is bounded and almost everywhere continuous in mean of order 2, it can be shown that the sums (16) will converge in probability to a random variable. But as in the case of the Cauchy integral this is technically an inconvenient definition for an integral. Naturally we attempt to repeat what we did with the sums

(15) and seek a limit for sums

$$\sum_{j=1}^{q} \phi(\tau_j) \, \Delta_j \tilde{z}^\rho \tag{18}$$

for all δ-fine partitions as σ tends to 0. But this limit is easily shown not to exist even for some very simple ϕ and $\tilde{z}^\rho$. Fortunately, only a small modification is needed to ensure that the limit in probability exists. Instead of using the δ-fine partitions defined in (4) and (5), we define a δ-fine belated partition to be a finite set $(t_0, t_1, \cdots, t_q, \tau_1, \cdots, \tau_q)$ such that

$$t_0 = a \le t_1 \le \cdots \le t_q = b \tag{19}$$

and

for $j = 1, \cdots, q_j$ is in the interval $[a', b']$ (that contains $[a,b]$)

and

$$\tau_j \le t_{j-1} \le t_j < \tau_j + \delta. \tag{20}$$

Now, if ϕ is bounded and almost everywhere continuous in second-order mean and is F_t-measurable for each t, the sums (16) for δ-fine belated partitions will converge in probability to a limit as δ tends to 0. This limit we call the "belated integral"

$$\int_a^b \phi(t) \, d\tilde{z}^\rho(t).$$

Still another of the sums is of the form

$$\sum_{j=1}^{q} \phi(t_{j-1}) \, \Delta_j \tilde{z}^\rho \, \Delta_j \tilde{z}^\sigma.$$

If we replace t_{j-1} by τ_j, where the τ_j satisfy (20), the sum becomes the partition-sum corresponding to a belated partition, and it can be shown that if ϕ is bounded and almost everywhere continuous in mean of first order the partition-sum will converge in probability to a random variable as δ tends to 0. This limit is called the second-order belated integral of ϕ with respect to $d\tilde{z}$ and $d\tilde{z}^\rho$, and is denoted by

$$\int_a^b \phi(t) \, d\tilde{z}^\rho \, d\tilde{z}^\sigma.$$

There are two other kinds of sums in the expression obtained by adding the right members of (14). One has factors $(\Delta_j t)^2$, the other has factors $\Delta_j t \, \Delta_j \tilde{z}^\rho$. Both these sums tend to 0 in probability as δ tends to 0.

Now all the integrals in the equation

$$x^i(t) = x^i(a) + \int_a^t f^i(\tau, x(\tau)) \, d\tau + \sum_{\rho=1}^{r} \int_a^t g^i{}_\rho(\tau, x(\tau)) \, dz^\rho(\tau)$$
$$+ \frac{1}{2} \sum_{\rho,\sigma=1}^{r} \int_a^t \Big\{ \sum_{k=1}^{n} (\partial g^i{}_\rho(\tau, x(\tau))/\partial x^k) g^k{}_\sigma(\tau, x(\tau)) \Big\} d\tilde{z}^\sigma(\tau) \, d\tilde{z}^\rho(\tau) \tag{21}$$

have meanings as Riemann or belated integrals. It does not follow instantly that the equations have solutions, nor that the $x_3{}^i(t_j)$ computed by (14) are good approximations to this solution (or, rather, than the solution of (21) is a good approximation to the $x_3(t_j)$, which are closer to physically significant quantities) when the partitioning is fine. This demands that a whole integral calculus be developed for the belated integrals, and also a theory of differential equations including equations (21). This is clearly not a task to be carried through in a few minutes, but it has been performed, and the results are available (1). Since the definitions of the integrals so closely resemble that of the Riemann integral, it is not surprising that the calculus of belated integrals closely resembles that of the Riemann integral, and in fact includes it (since the case of a "stochastic process" in which one single z has probability 1 is merely the deterministic case).

Equations (21) are said to be "in canonical form", and the integrand in the last integral is abbreviated to

$$g^i_{\rho,\sigma} = \sum_{k=1}^{n} (\partial g^i_\rho / \partial x^k) g^k_\sigma \; .$$

Such equations have a large number of useful properties not shared by other equations involving stochastic integrals; we do not have time even to list those properties. For our present purposes, the crucial point is that when $r = 1$, or more generally when

$$g^i_{\rho,\sigma} = g^i_{\sigma,\rho} \quad (i=1,\cdots,n;\ \rho,\sigma = 1,\cdots,r),$$

the approximations $x^i{}_3$ to the solution of (2) with Lipschitzian noises, and the similar approximations $\tilde{x}_3$ to the solutions of equations (21) with noises $\tilde{z}^\rho$ that satisfy (17), converge rapidly to the solution of the differential equations, hence that the solutions of the differential equations are acceptably close to x_3 and $\tilde{x}_3$ when the intervals $[t_{j-1},t_j]$ are long enough so that the distribution of $(z^\rho(t_0),\cdots,z^\rho(t_q))$ is well approximated by that of $(\tilde{z}^\rho(t_0),\cdots,\tilde{z}^\rho(t_q))$.

III. SLOWLY RESPONDING SYSTEMS. Having made all these complimentary (or self-gratulatory) remarks about the canonical form of stochastic differential equations, we now shall look at an example in which the canonical form is inappropriate. A good example of this type is furnished by the stock market, as discussed by Barrett and Wright (2). Here the disturbances are sales of securities, and the state variables x^i are indices that specify the state of the market. Since each z^ρ is the total of sales of something up to time t, it is constant between jumps, and an integral with respect to z^ρ is in fact a finite sum. A change in z produces a change in the x^i, but there is a delay t* in publicizing this change in the x^i. So the sensitivity g^i of x^i to change in z is a function of the values of the x^i at the time $t - t^*$. Thus the x^i, instead of satisfying a differential equation (2), will satisfy an equation of the form

$$\begin{aligned} x^i(t) = x^i(a) &+ \int_a^t f^i(\tau-t^*,\ x(\tau-t^*))\, d\tau \\ &+ \sum_{\rho=1}^{r} \int_\sigma^\tau g^i_\rho(\tau-t^*,\ x(\tau-t^*))\, dz^\rho(\tau). \end{aligned} \tag{22}$$

To estimate this by a finite sum we first subdivide [a,b] by points $t_0 = a < t_1 < \cdots < t_q = b$ with all $t_j - t_{j-1}$ less than t*. We again use the modified Euler method, replacing each integrand in (22) by the arithemetic mean of its values at $\tau = t_{j-1} - t^*$ and at $\tau = t_j - t^*$. Then the right member of (22) is approximately

$$\begin{aligned} x^i(a) &+ \frac{1}{2}\sum_{j=1}^{q} f^f(t_{j-1}-t^*,\ x(t_{j-1}-t^*))\,\Delta_j t \\ &+ \frac{1}{2}\sum_{j=1}^{q} f^i(t_j-t^*,\ x(t_j-t^*))\,\Delta_j t \\ &+ \frac{1}{2}\sum_{j=1}^{q}\sum_{\rho=1}^{r} g^i{}_\rho(t_{j-1}-t^*,\ x(t_{j-1}-t^*))\,\Delta_j z^\rho \\ &+ \frac{1}{2}\sum_{j=1}^{q}\sum_{\rho=1}^{r} g^i{}_\rho(t_j-t^*,\ x(t_j-t^*))\,\Delta_j z^\rho. \end{aligned} \tag{23}$$

But the set of numbers

$$(t_0, t_1, \cdots, t_q,\ t_0-t^*, \cdots, t_{q-1}-t^*) \tag{24}$$

is a belated partition of [a,b], and so is the set

$$(t_0, t_1, \cdots, t_q,\ t_1-t^*, \cdots, t_q-t^*). \tag{25}$$

The first and third sums in (23) are partition-sums corresponding to the belated partition (24), and the second and fourth are partition-sums corresponding to the belated partition (25). So if t* is small (23) is a good approximation to

$$x^i(a) + \int_a^b f^i(\tau-t^*,\ x(\tau-t^*))\,d\tau + \sum_{\rho=1}^{r}\int_a^b g^i{}_\rho(\tau-t^*,\ x(\tau-t^*))\,dz^\rho(\tau). \tag{26}$$

No second-order integrals are needed. Moreover, if t* is small the quantity (26) will be near the value at b of the solution x of the equation

$$x^i(t) = x^i(a) + \int_a^t f^i(\tau,\ x(\tau))\,d\tau + \sum_{\rho=1}^{r}\int_a^t g^i{}_\rho(\tau,\ x(\tau))\,dz^\rho(\tau). \tag{27}$$

The last integral is a first-order belated or Itô integral. The delay t* makes the model different from the first example by the absence of second-order integrals.

IV. STOCHASTIC OPTIMAL CONTROL. For a final and most complicated example we consider a problem in stochastic optimal control. We suppose that functions f^i and g^i are defined for all $(t, x^1, \cdots, x^n)$ and all u in a set U in a finite-dimensional space. If the z are Lipschitzian functions, satisfying (3), and $t \mapsto u(t)$ is a sufficiently well-behaved function with values in U, the equations

$$\begin{aligned} x^i(t) = x^i(a) &+ \int_a^t f^i(\tau,\ x(\tau),\ u(\tau))\,d\tau \\ &+ \sum_{\rho=1}^{r}\int_a^t g^i{}_\rho(\tau,\ x(\tau),\ u(\tau))\,dz^\rho(\tau) \end{aligned} \tag{28}$$

will be solvable on [a,b]. We wish to choose a function u, depending on t and on available information about x(t), for which a certain cost function C(x(b)) will have the smallest expectation among all such functions u. Even in the absence of noises, non-linear control problems are very difficult, and stochastic control problems cannot be any easier. However, by ill-advised idealization we can make the problem much more difficult, and even obscure its very meaning. The purpose of an idealization is to replace the problem by a more tractable substitute. If our idealization makes things harder, it is a failure and should be replaced by something better.

In (28) it is tempting to replace u by u(t,x(t)), where u(t,x) has values in U. But this is unrealistic. It would assume a knowledge of x(t) for every t in [a,b], which would involve knowing infinitely many bits of information. We are closer to reality if we assume that information about x(t) is acquired at the times t_j in a set $\Pi = \{t_0,\cdots,t_q\}$. If at each time t we know the values of $x(t_j)$ at those times t_j in Π such that $a \leq t_j \leq t$, then on the interval $(t_{j-1},t_j]$ we acquire no more information about x. We choose a function u on $(t_{j-1}, t_j]$ which depends on the known values $x(t_0,),\cdots,x(t_{j-1})$ but not on other values of x. For each such choice of a function u equations (28) are of the same type as equations (2). If we wish to idealize the noise-system by replacing the z^ρ by some other processes satisfying conditions (17), as before we re-write the differential equations in canonical form. The problem now becomes a succession of "open-loop" problems. Such problems have been discussed by V. M. Warfield (3) under the assumption that the increments of the z^ρ in different intervals $[t_{j-1},t_j]$ are independent. Let C_q be the function we called C, so that the expectation of the cost is the expectation of $C_q(x(t_q)) = C_q(x(b))$. To each point $x = (x^1,\cdots,x^n)$ in n-space there corresponds a function u on $[t_{q-1},t_q]$ that minimizes the expectation of $C_q(x(t_q))$ under the condition that $x(t_{q-1}) = x$. This minimum value of the expected cost is a function of x, which we call $C_{q-1}(x)$. For each point x in n-space there is a function u on $[t_{q-2},t_{q-1}]$ that minimizes the expectation of $C_{q-1}(x(t_{q-1}))$ under the condition $x(t_{q-2}) = x$. Continuing backwards, u is determined on each interval $[t_{j-1},t_j]$ as a function of t and $x(t_{j-1})$.

The action of the disturbances z^ρ has now been idealized from a finite-sum procedure to an *integral*. But we are still left with a finite step-process, depending on the points in the set Π of times of acquiring information. It is a reasonable conjecture, verifiable in at least one example, that this finite step process can profitably be replaced by something in the nature of its limit as the length of the longest interval $[t_{j-1},t_j]$ tends to 0, just as finite partition-sums can profitably be replaced by the integrals to which they converge. When this is possible, it is that limit that we shall accept as the idealization of the physically unattainable concept of "optimal control in the presence of continuously acquired information".

Let X be the class of continuous functions $(x^1(t),\cdots,x^n(t))$ on [a,b], and let V be the class of functions $(t,x(\cdot)) \mapsto u(t,x(\cdot))$ with values in U defined for t in [a,b] and for $x(\cdot)$ a function belonging to the class X, and almost everywhere continuous in t for fixed choice of $x(\cdot)$. To each set $\Pi = \{t_0,\cdots,t_q\}$ with $t_0 = a < t_1 <\cdots< t_q = b$ there corresponds a subset V_Π of V consisting of those u which on each interval $(t_{j-1},t_j]$ depend only on t and $x(t_{j-1})$. If the z^ρ satisfy (17), and in (28) we replace $u(\tau)$ by $u(\tau,x(\cdot))$ for any member $u(t,x(\cdot))$ of V_Π, equations (28) will have a solution, and the expectation of C(x(b)) for that solution will be a real-valued function on V_Π. Let its greatest lower bound be denoted by $\mu(\Pi)$. This is then the greatest lower bound of the expected cost when the choice of control at time t is based only on knowledge of the values of the $x^i(t)$ at times $t_0,\cdots,t_{j-1}$ in Π preceding t. It may happen that there is a function $(t,x) \mapsto u^*(t,x)$, defined for all t in [a,b] and all real $x^1,\cdots,x^n$ and with values in U, such that if for each set Π we choose

$$u(t,x(\cdot)) = u^*(t,x(t_{j-1})) \quad (t_{j-1} < t \leq t_j), \tag{29}$$

the solution of (28) with this control will be nearly optimal, in the sense that for each positive ϵ there is a positive δ such that whenever Π is a set with max $(t_j-t_{j-1}) < \delta$, the solution of (28) with control (29) will give an expected cost less than $\mu(\Pi) + \epsilon$. In this case u* will idealize the concept of "optimal control in the presence of continuously acquired information". It does not really demand knowledge of x(t) for all t, any more than the definition of the integral requires that all intervals $[t_{j-1},t_j]$ have length 0. Instead, it is a limit of quantities each based on a finite set, and thus conceivably attainable.

I am not aware of any published attempt to carry out this program in any specific example, and it remains pure conjecture that it would be helpful in any complicated situation. However, in one simple example it can be carried through in detail, and it clarifies a problem whose meaning (let alone its solution) is not easily comprehended otherwise.

Let U be the interval [-1,1], and let z be a process that satisfies (17) and has increments $z(t) - z(s)$ $(0 \leq s < t \leq 1)$ that are symmetrically distributed about 0 and are independent over disjoint intervals. The state variable x satisfies

$$x(t) = x(0) + \int_0^t u \, d\tau + \int_0^t dz(\tau). \tag{30}$$

We seek to minimize the expectation of $x(1)^2$. It is easy to conjecture a rule for minimizing that expectation; the rule is "at all times t at which $x(t) > 0$, choose $u(t) = -1$; at all times t at which $x(t) < 0$, choose $u(t) = +1$". But if z is a Wiener process, if it vanishes at time t' it will change sign infinitely often in every open interval that contains t'. We cannot even solve equations (28); and if we could, the requirement to change u from -1 to +1 and back again infinitely often in a nanosecond cannot be met. Moreover, we would need to know the exact value of x(t) at all times t in [a,b], which again asks us to know infinitely many bits. Instead of this conjecture, we need a solution that can be applied when x(t) is known at the times t in a finite set $\Pi = \{t_0,\cdots,t_q\}$, and that gives nearly the least possible value to $E(x(b)^2)$ when the intervals $[t_{j-1},t_j]$ are all short.

If for t in [0,1] we know the values of $x(t_j)$ at all times t_j in Π that are not later than t, the choice of x(t) on interval $[t_{j-1},t_j]$ is to be made on the basis of that information. That is, the permitted controls have values on $[t_{j-1},t_j]$ that depend on t and on $x(t_0),\cdots,x(t_{j-1})$ only. Among all such functions the optimal function u_Π can be specifically determined. Let $\tilde{t}_j = \min\{t_j, t_{j-1} + |x(t_{j-1})|\}$. On $[t_{j-1},\tilde{t}_j)$ choose

$$u_\Pi(t) = +1 \text{ if } x(t_{j-1}) < 0,$$

$$= -1 \text{ if } x(t_{j-1}) > 0;$$

on $[\tilde{t}_j,t_j)$ choose $u_\Pi(t) = 0$. Corresponding to this u equation (28) has a solution which we call x_Π, and we can show that $E(x_\Pi(b)^2)$ is the least value of the expected cost for all permitted controls.

Now let us define

$$u^*(t,x) = 1 \quad (x < 0)$$
$$= 0 \quad (x = 0) \tag{31}$$
$$= -1 \quad (x > 0).$$

If for each $\Pi = \{t_0, \cdots, t_q\}$ we use the control which on $[t_{j-1}, t_j)$ is constantly equal to $u^*(t_{j-1}, x(t_{j-1}))$, we obtain a solution x^* of equation (28) such that the expectation of $x^*(b)^2$ is not exactly equal to the optimal value $E(x_\Pi(b)^2)$ given by the optimal control u_Π. But the amount by which $E(x^*(b)^2)$ exceeds the least possible value $E(x_\Pi(b)^2)$ tends to 0 with max $(t_j - t_{j-1})$. The application of u^* to any Π is the simple "bang bang" rule: whenever you find the value of $x(t_{j-1})$ for some time t_{j-1}, set u at +1 if $x(t_{j-1}) < 0$ and at -1 if $x(t_{j-1}) > 0$, and leave u at that value until the next determination of x comes in. If information about $x(t)$ comes in frequently, the result will be nearly the best possible.

Similar considerations should apply to other problems, but while the conceptual troubles are thereby resolved, the computational troubles remain with us.

REFERENCES

(1) E. J. McShane, Stochastic calculus and stochastic models, Academic Press 1974.

(2) J. F. Barrett and D. J. Wright, The random nature of stockmarket prices, Operations Research 22 (1974), pp. 175-177.

(3) Virginia M. Warfield, A stochastic maximum principle, J. Control and Optimization 14 (1976), pp. 803-826.

SOME BOUNDS FOR OPTIMAL MANEUVERS AND PREDICTORS

Harry L. Reed, Jr.

US Army Armament Research and Development Command
Ballistic Research Laboratory
Aberdeen Proving Ground, MD 21005

ABSTRACT

A problem of some interest in the understanding of fire control systems is the following pure prediction problem.

Given the class of target trajectories $\underline{\overline{X}}$ for which

$$\lim_{B \to \infty} \frac{1}{2B} \int_{-B}^{B} [\ddot{x}(t)]^2 dt = a^2 \quad \text{where } x \in \underline{\overline{X}}$$

and the class of predictors P which satisfy the causal principle, find

$$\varepsilon_1 = \sup_{x \in \underline{\overline{X}}} \; \inf_{p \in P} \varepsilon$$

and

$$\varepsilon_2 = \inf_{p \in P} \; \sup_{x \in \underline{\overline{X}}} \varepsilon$$

where

$$\varepsilon^2 = \frac{1}{2B} \int_{-B}^{B} [x(t+T) - y(t+T)]^2 dt$$

and $y(t+T)$ is the predicted value of $x(t+T)$ based on values of $x(t-\tau)$ with $\tau \geqslant 0$.

It is shown that

$$\varepsilon_1 = \varepsilon_2$$

for general N. For the particular value of $N = 2$ which corresponds to a limited acceleration for the target, we have

$$\varepsilon_1 = \varepsilon_2 = \frac{2}{(\lambda T)^2} [\tfrac{1}{2} a T^2] \cong 0.569 [\tfrac{1}{2} a T^2]$$

where λ is determined from the solution of an eigenvalue problem for a fourth-order differential equation.

The prediction algorithm p_α for which

$$\varepsilon_1 = \varepsilon(x_\alpha, p_\alpha)$$

is a linear operator and the optimal subclass of maneuvers x_α is based on a second-order correlation function

$$E[\ddot{x}(t)\ddot{x}(t+\tau)] = \int_0^\infty \alpha(s)\alpha(s+\tau)ds.$$

For the particular case of $N = 2$ we have

$$\alpha(s) = \frac{a}{\sqrt{T}} \left[\frac{\cosh \lambda(s - T/2)}{\cosh(\lambda T/2)} - \frac{\sin \lambda(s - T/2)}{\sin \lambda T/2} \right]$$

for

$$0 \leqslant s \leqslant T$$

and

$$\alpha(s) = 0 \quad \text{otherwise.}$$

No restrictions were placed on $\underline{X}$ and P other than those stated above (i.e., $\underline{X}$ was not restricted to stationary processes and P was not restricted to linear operators).

It is further shown that the strategies given above are good approximations for the more general analysis in which hit probability is the performance measure. This is the case at least for "first cut" analyses.

Bounds such as this help avoid the expenditure of resources to achieve the impossible or to achieve marginally small improvements in fire control design.

1. INTRODUCTION

The goal of this paper is to give some insights into how well a fire control system can be expected to perform with noiseless information and how well a target can avoid being hit with limits on its ability to maneuver. Such a goal is very ambitious, so we shall make three simplifying assumptions:

- The tracking data are noiseless. This gives an advantage to the gun (see Reference 1) but is not too significant since optical and millimeter radar systems promise very accurate tracking and also since in many cases the errors from evasive maneuvers far exceed errors resulting from errors in state estimation.

- The target is limited only in the r.m.s. value of the N^{th} derivative of its path. That is, the class of maneuvers $\underline{\overline{X}}_N$ is limited to those $x(t)$ for which

$$C_N{}^2 = \lim_{R \to \infty} \frac{1}{2R} \int_{-R}^{R} [x^{(N)}(t)]^2 dt. \qquad 1.1$$

- The performance of the fire control is characterized by the r.m.s. prediction error

$$\varepsilon^2(x,p) = \lim_{R \to \infty} \frac{1}{2R} \int_{-R}^{R} [\hat{x}(t+T,p) - x(t+T)]^2 dt \qquad 1.2$$

where T is the time of flight of the bullet, $p \in P$, P is the class of prediction algorithm such that $\hat{x}$ is the predicted value of $x(t+T)$ given all data on $x(t-s)$ for $s \geq 0$.

The last two assumptions are the closest concession we make to stationarity. The reason for averaging over time in Equation 1.2 is to provide a measure that does not encourage the target to make a one-time maneuver at the time of firing a single round but rather forces the target to avoid rounds fired at unknown times or to avoid bursts of rounds fired over time.

[1] Harry L. Reed, Jr., "Some Bounds on the Generalized Fire Control Problem," Ballistic Research Laboratories Report No. 1946, November 1976 (AD A033043).

Finally, the use of an r.m.s. error gives an incomplete measure of effectiveness for maneuvers with statistics that do not allow an adequate measure of probability of hit from the r.m.s. error (see, for example, Reference 2). However, for optimal maneuvers, the r.m.s. error is a fairly good measure (see Section 5).

We shall omit the subscript N unless its particular value is important. The following is our overall strategy:

Let
$$\varepsilon_o = \sup_{x \in \overline{\underline{X}}} \; \inf_{p \in P} \varepsilon(x,p) \qquad 1.3$$

$$\varepsilon^o = \inf_{p \in P} \; \sup_{x \in \overline{\underline{X}}} \varepsilon(x,p) \qquad 1.4$$

$$\tilde{\varepsilon}_o = \varepsilon(x_o,p_o) = \sup_{x \in \overline{\underline{X}}_G} \; \inf_{p \in P} \varepsilon(x,p) \qquad 1.5$$

$$\tilde{\varepsilon}^o = \sup_{x \in \overline{\underline{X}}} \varepsilon(x,p_o) \qquad 1.6$$

where $\overline{\underline{X}}_G$ is the class of stationary Gaussian maneuvers that satisfy Equation 1.1 and also satisfy

$$\int_{-\infty}^{\infty} \frac{|\log|\Phi(w)||}{1 + w^2} \, dw < \infty \qquad 1.7$$

where $\Phi(w)$ is the power spectral density of the Nth derivative.

Since $\overline{\underline{X}}_G \subset \overline{\underline{X}}$ and $p_o \in P$, we have

$$\tilde{\varepsilon}_o \leq \varepsilon_o \leq \varepsilon^o \leq \tilde{\varepsilon}^o \qquad 1.8$$

The argument that gives the middle inequality is a consequence of the properties of sup and of inf, is common knowledge in game theory, and is given in Appendix 1 for the benefit of the uninitiated.

We shall show that

$$\tilde{\varepsilon}_o = \tilde{\varepsilon}^o \qquad 1.9$$

and thus that

$$\varepsilon_o = \varepsilon^o . \qquad 1.10$$

[2] Harry L. Reed, Jr., " Limitations of the R.M.S. Criterion for Fire Control," Ballistic Research Laboratories Report No. 1805, July 1975 (AD A014986).

We shall also evaluate ε_o for N = 0, 1, and 2 and show how to evaluate it for higher values of N.

Equation 1.10 implies that in a game between two "smart" players, x_o and p_o are optimal strategies.

2. LOWER BOUND

For Gaussian maneuvers the optimal predictors are linear operators of the form (see Reference 3)

$$\hat{x}(t+T,p_h) = \sum_{m=o}^{N-1} x^{(m)}(t)\,\frac{T^m}{m!} + \int_o^\infty h(s)x^{(N)}(t-s)ds \qquad 2.1$$

Integration by parts gives

$$\begin{aligned} x(t+T) - \hat{x}(t+T,\, p_h) &= \int_o^\infty u_N(T-s)x^{(N)}(t+s)ds \\ &\quad - \int_o^\infty h(s)x^{(N)}(t-s)ds \end{aligned} \qquad 2.2$$

where

$$u_o(t) = \delta(t) \qquad 2.3$$

and for N > 0

$$u_N(t) = \frac{t^{N-1}}{(N-1)!} \qquad \text{for } t \geqslant 0 \qquad 2.4$$

$$= 0 \qquad \text{for } t < 0. \qquad 2.5$$

Again we shall use

$f(t)$ for $u_N(T-t)$

and

$a(t)$ for $x^{(N)}(t)$

unless the particular value of N is important to the argument at hand.

[3] Norbert Wiener, *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*, The Technology Press of M.I.T. and John Wiley & Sons, Inc., New York.

Let

$$\phi(s) = \lim_{R \to \infty} \frac{1}{2R} \int_{-R}^{R} a(t)a(t+s)dt. \qquad 2.6$$

Equations 1.1 and 1.7 allow us to write

$$\phi(s) = \int_{-\infty}^{\infty} \alpha(t)\alpha(t+s)dt \qquad 2.7$$

where

$$\alpha(t) = 0 \qquad t < 0 \qquad 2.8$$

and of course

$$C^2 = \int_{0}^{\infty} [\alpha(t)]^2 dt. \qquad 2.9$$

Using

$$\phi(r-s) = \int_{-\alpha}^{\alpha} \alpha(t+r)\alpha(t+s)dt \qquad 2.10$$

$$= \int_{-\alpha}^{\alpha} \alpha(t-r)\alpha(t-s)dt \qquad 2.11$$

and

$$\phi(r+s) = \int_{-\alpha}^{\alpha} \alpha(t-r)\alpha(t+s)dt, \qquad 2.12$$

we can combine Equation 1.2 and 2.2 to write

$$\varepsilon^2 = \int_{-\infty}^{\infty} dt \left\{ \int_{0}^{\infty} [f(s)\alpha(t+s) - h(s)\alpha(t-s)]ds \right\}^2 \qquad 2.13$$

$$= \int_{0}^{\infty} dt \left\{ \int_{0}^{\infty} [f(s)\alpha(t+s) - h(s)\alpha(t-s)]ds \right\}^2$$

$$+ \int_{-\infty}^{0} dt \left\{ \int_{0}^{\infty} f(s)\alpha(t+s)ds \right\}^2 . \qquad 2.14$$

To minimize with respect to p, we pick h to satisfy

$$\int_0^\infty f(s)\alpha(t+s)ds = \int_0^\infty h(s)\alpha(t-s)ds \qquad 2.15$$

which puts the first term of the function in Equation 2.14 equal to zero.

To maximize with respect to x, we then pick α to maximize

$$\begin{aligned} \varepsilon^2 &= \int_{-\infty}^{0} dt \left\{ \int_0^\infty f(s)\alpha(t+s)ds \right\}^2 \\ &= \int_0^\infty dt \left\{ \int_0^\infty f(s)\alpha(s-t)ds \right\}^2 \qquad 2.16 \\ &= \int_0^\infty dt \left\{ \int_0^\infty f(s+t)\alpha(s)ds \right\}^2 . \end{aligned}$$

To do this, we set

$$\delta\varepsilon^2 = 2\int_0^\infty dt \int_0^\infty f(s+t)\alpha(s)ds \int_0^\infty f(r+t)\delta\alpha(r)dr = 0 \qquad 2.17$$

subject to

$$\int_0^\infty \alpha(r)\ \delta\alpha(r) = 0. \qquad 2.18$$

Therefore

$$\alpha(r) = k\int_0^\infty dt \int_0^\infty f(r+t)\ f(s+t)\alpha(s)ds. \qquad 2.19$$

Multiplying Equation 2.19 by α(r) and integrating, we have

$$\varepsilon^2 = \frac{C^2}{k} \qquad 2.20$$

which shows that $k \geqslant 0$. Further

$$\varepsilon_o{}^2 = \frac{C^2}{k_o}$$

where k_o is the least eigenvalue of Equations 2.9 and 2.19.

In Section 4 we show how this eigenvalue problem is related to an eigenvalue problem for a system of differential equations and we evaluate k_o for N = 0, 1, and 2.

3. UPPER BOUND

Even though the class $\underline{X}$ is only constrained by

$$C^2 = \lim_{R \to \infty} \frac{1}{2R} \int_{-R}^{R} [a(t)]^2 dt$$

we can define

$$\phi(s) = \lim_{R \to \infty} \frac{1}{2R} \int_{-R}^{R} a(t)a(t+s)dt \qquad 3.1$$

and know that

$$\Phi(w) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \phi(s)e^{-iws}ds \geq 0. \qquad 3.2$$

Now we shall use the predicter which was defined in the previous section by Equation 2.15 for the particular $\alpha(r)$ given in Equation 2.19. Then Equations 1.7, 2.2, and 3.1 give (for any x)

$$\varepsilon^2(x,p_o) = \int_0^{\infty} dr \int_0^{\infty} ds \Big\{ f(s)f(r)\phi(r-s) - 2\, f(s)h(r)\phi(r+s) + h(r)h(s)\phi(r-s) \Big\}. \qquad 3.3$$

Using

$$\phi(r) = \int_{-\infty}^{\infty} \Phi(w)e^{iwr}dw, \qquad 3.4$$

we can derive

$$\varepsilon^2(x,p_o) = \int_{-\infty}^{\infty} \mid \overline{F}(w) - H(w) \mid^2 \Phi(w)dw \qquad 3.5$$

where

$$H(w) = \int_0^\infty h(t)e^{-iwt}dt \qquad 3.6$$

and

$$F(w) = \int_0^\infty f(t)e^{-iwt}dt \qquad 3.7$$

We shall show that

$$|\overline{F}(w) - H(w)|^2 = \frac{1}{k_o} \qquad 3.8$$

and thus that

$$\varepsilon^2(x,p_o) = \frac{1}{k_o}\int_{-\infty}^\infty \Phi(w)dw = \frac{C^2}{k_o} \qquad 3.9$$

for all $x \in \underline{\overline{X}}$.

To do that, we first take the Fourier transform of Equation 2.15 to get

$$\int_0^\infty e^{-iwt}dt \int_0^\infty f(s)\alpha(t+s)ds = H(w)A(w) \qquad 3.10$$

where

$$A(w) = \int_0^\infty \alpha(t)e^{-iwt}dt \qquad 3.11$$

Some manipulation of the left hand side of Equation 3.10 gives

$$\begin{aligned}
&\int_0^\infty e^{-iwt}dt \int_0^\infty f(s)\alpha(t+s)ds \\
&= \int_{-\infty}^\infty e^{-iwt}dt \int_0^\infty f(s)\alpha(t+s)ds - \int_{-\infty}^0 e^{-iwt}dt \int_0^\infty f(s)\alpha(t+s)ds \\
&= \overline{F}(w)A(w) - \int_0^\infty e^{iwt} \int_0^\infty f(s)\alpha(s-t)ds \\
&= \overline{F}(w)A(w) - \int_0^\infty e^{iwt} \int_0^\infty f(t+r)\alpha(r)dr
\end{aligned} \qquad 3.12$$

Therefore

$$|\overline{F}(w) - H(W)|^2 = |B(w)/A(w)|^2 \qquad 3.13$$

where

$$B(w) = \int_{-\infty}^{\infty} e^{iwt}\beta(t)dt \qquad 3.14$$

$$\beta(t) = \int_0^{\infty} f(t+r)\alpha(r)rt \quad t > 0 \qquad 3.15$$

$$= 0 \quad t < 0$$

We then have

$$|B(w)|^2 = B(w)\overline{B}(w) = \int_{-\infty}^{\infty} e^{iwp}dp \int_0^{\infty} \beta(q)\beta(p+q)dq \qquad 3.16$$

Note that this convolution is an even function of p so that

$$|B(w)|^2 = \int_{-\infty}^{\infty} e^{iwp}dp \int_0^{\infty} \beta(q)\beta(|p|+q)dq \qquad 3.17$$

Now using Equation 2.19, we have

$$\begin{aligned} \int_0^{\infty} \beta(q)\beta(|p|+q)dq &= \int_0^{\infty} dq \int_0^{\infty} f(q+r)\alpha(r)dr \\ &\times \int_0^{\infty} f(q+|p|+s)\alpha(s)ds \\ &= \frac{1}{K_0}\int_0^{\infty} \alpha(|p|+s)\alpha(s)ds \end{aligned} \qquad 3.18$$

and finally

$$|B(w)\overline{B}(w)| = \frac{1}{k_o}\int_{-\infty}^{\infty} e^{iwp}dp \int_0^{\infty} \alpha(|p|+s)\alpha(s)ds$$

$$= \frac{1}{k_o}\ |A(w)\overline{A}(w)| \qquad 3.19$$

and we have from Equation 3.13 and 3.5 that

$$\varepsilon^2(x,p_o) = \frac{C^2}{k_o}$$

as advertised in Equation 3.9.

Another approach to Equation 3.8 is to show that

$$B(w) = \frac{1}{\sqrt{k_o}}\ \overline{A}(w) \qquad 3.20$$

which can be shown by showing that

$$\beta(t) = \frac{1}{\sqrt{k_o}}\ \alpha(t). \qquad 3.21$$

To do this, we combine Equation 2.19 and 3.15 to get

$$\beta(t) = \int_0^{\infty} f(t+r)dr \int_0^{\infty} k_o dq \int_0^{\infty} f(r+q)f(s+q)\alpha(s)ds$$

$$= k_o\int_0^{\infty} f(t+r)dr \int_0^{\infty} f(r+q)\beta(q)dq \qquad 3.22$$

Thus $\beta(t)$ satisfies the same integral equation as $\alpha(t)$. In the next section we relate this integral equation to the eigenvalue problem for a differential equation. This eigenvalue problem has only one linearly independent solution (see Appendix 2) and so we can write

$$\beta(t) = \gamma\ \alpha(t) \qquad 3.23$$

and finally

$$|B(w)\overline{B}(w)| = \frac{1}{k_o}\int_{-\infty}^{\infty} e^{iwp}dp \int_o^{\infty} \alpha(|p|+s)\alpha(s)ds$$

$$= \frac{1}{k_o}\,|A(w)\overline{A}(w)| \qquad 3.19$$

and we have from Equation 3.13 and 3.5 that

$$\varepsilon^2(x,p_o) = \frac{C^2}{k_o}$$

as advertised in Equation 3.9.

Another approach to Equation 3.8 is to show that

$$B(w) = \frac{1}{\sqrt{k_o}}\,\overline{A}(w) \qquad 3.20$$

which can be shown by showing that

$$\beta(t) = \frac{1}{\sqrt{k_o}}\,\alpha(t). \qquad 3.21$$

To do this, we combine Equation 2.19 and 3.15 to get

$$\beta(t) = \int_o^{\infty} f(t+r)dr \int_o^{\infty} k_o dq \int_o^{\infty} f(r+q)f(s+q)\alpha(s)ds$$

$$= k_o\int_o^{\infty} f(t+r)dr \int_o^{\infty} f(r+q)\beta(q)dq \qquad 3.22$$

Thus $\beta(t)$ satisfies the same integral equation as $\alpha(t)$. In the next section we relate this integral equation to the eigenvalue problem for a differential equation. This eigenvalue problem has only one linearly independent solution (see Appendix 2) and so we can write

$$\beta(t) = \gamma\,\alpha(t) \qquad 3.23$$

Then

$$\int_o^\infty [\beta(t)]^2 dt = \gamma^2 \int_o^\infty [\alpha(t)]^2 dt \qquad 3.24$$

which gives

$$\int_o^\infty \left\{ \int_o^\infty f(t+r)\alpha(r)dr \right\}^2 dt = \gamma^2 \int_o^\infty [\alpha(t)]^2 dt \qquad 3.25$$

using the definition of $\beta(t)$ and which can be rewritten as

$$\int_o^\infty \alpha(r)dr \int_o^\infty f(t+r)dt \int_o^\infty f(t+s)\alpha(s)ds$$

$$= \gamma^2 \int_o^\infty [\alpha(t)]^2 dt \qquad 3.26$$

and finally (from Equation 2.19)

$$\frac{1}{k_o} \int_o^\infty [\alpha(r)]^2 dr = \gamma^2 \int_o^\infty [\alpha(t))]^2 dt \qquad 3.27$$

which gives

$$\gamma = \frac{1}{\sqrt{k_o}} \qquad 3.28$$

4. THE EIGENVALUE PROBLEM

We have

$$\int_o^\infty \alpha^2(r) = C^2 \qquad 4.1$$

and

$$\alpha(r) = k_o \int_o^\infty dt \int_o^\infty u(T-r-t)u(T-s-t)\alpha(s)ds$$

$$= k_o \int_o^{T-r} dt \int_o^{T-t} \frac{(T-r-t)^{N-1}}{(N-1)!} \frac{(T-s-t)^{N-1}}{(N-1)!} \alpha(s)ds \qquad 4.2$$

for $r < T$.

We also have $k_o > 0$, and k_o is the least eigenvalue of this system of equations.

If $r > T$

$$u(T-r-t) = 0 \quad \text{since} \quad t \geq 0$$

and thus

$$\alpha(r) = 0 \quad \text{for} \quad r > T \qquad 4.3$$

If $N = 0$

$$u(r) = \delta(r)$$

and

$$\alpha(r) = k_o \int_0^\infty dt \int_0^\infty \delta(T-r-t)\delta(T-s-t)\alpha(s)ds$$

$$= k_o \; \alpha(r)$$

So for $N = 0$

$$k_o = 1 \qquad 4.4$$

and

$$\varepsilon_o = C \qquad 4.5$$

Let $N \geq 1$. We can differentiate Equation 4.2 to get for $0 \leq t \leq T$

$$\alpha^{(2N)} = (-1)^N k_o \alpha \qquad 4.6$$

$$\alpha^{(M)}(T) = 0 \quad \text{for} \quad M=0 \text{ to } N-1 \qquad 4.7$$

$$\alpha^{(M)}(0) = 0 \quad \text{for} \quad M=N \text{ to } 2N-1 \qquad 4.8$$

The uniqueness of the solution to this system of equations is shown in Appendix 2.

Now let N=1

$$\ddot{\alpha} = - k_o \alpha$$

$$\dot{\alpha}(o) = \alpha(T) = 0$$

$$\alpha = \sqrt{2/T}\ C \cos\left(\frac{\Pi}{2T}\, t\right) \qquad 4.9$$

$$k_o = \left(\frac{\Pi}{2T}\right)^2 \qquad 4.10$$

$$\varepsilon_o = \frac{2}{\Pi}\, CT \qquad 4.11$$

Finally let N = 2

$$\ddddot{\alpha} = k_o\alpha$$

$$\dddot{\alpha}\ (o) = \ddot{\alpha}(o) = \dot{\alpha}(T) = \alpha(T) = 0$$

(The classical problem of the vibration of a clamped rod.)

Letting $k_o = \lambda_o{}^4$, we have

$$1 + \cosh\ (\lambda_o T)\ \cos\ (\lambda_o T) = 0 \qquad 4.12$$

$$\lambda_o T \cong 1.875$$

$$\alpha = \frac{C}{\sqrt{T}} \left\{ \frac{\cosh[\lambda_o(t-T/2)]}{\cosh[\lambda_o T/2]} - \frac{\sin[\lambda_o(t-T/2)]}{\sin[\lambda_o T/2]} \right\}$$

$$\varepsilon_o = \frac{2}{\left(\lambda_o T\right)^2} \left(\frac{1}{2}\, CT^2\right) \qquad 4.14$$

$$\cong .569 \left(\frac{1}{2}\, CT^2\right) \qquad 4.15$$

5. THE GENERAL PROBLEM

In this section we shall see how far we can go using hit probability rather than the r.m.s. criterion. In doing this, we shall have to give up the neatness of finding an exact answer. On the other hand, we shall find bounds on the problem, and these bounds will be shown to bracket the problem closely enough for many "first analyses."

Let us discuss the problem where x(t) has a single spatial dimension. Associated with a class of maneuvers is a probability distribution function

$$u\{x(t+T) - \hat{x}(t+T,p) \mid x(t-s),\ s \geq 0\}. \qquad 5.1$$

That is, u is the distribution of the error between the future position and the predicted future position given the past. Let y be this error in future position. We can average over time to find a distribution function

$$u(y \mid x,p). \qquad 5.2$$

The probability of hit q is

$$q(x,p) = \int_{-\ell/2}^{\ell/2} du(\xi \mid x,p), \qquad 5.3$$

where ℓ is the size of the target. The pilot wishes to keep q small. His goal might be

$$q^o = \inf_{x \in X} \ \sup_{p \in P} \ q(x,p) \cdot \qquad 5.4$$

Likewise, the gunner might try for

$$q_o = \sup_{p \in P} \ \inf_{x \in X} \ q(x,p) \cdot \qquad 5.5$$

The set $\underline{X}$, the set P, the set X_G, the maneuver $\underline{X}_O$, the prediction algorithm p_o, and the error ε_o are as defined in Section 1.

We can define

$$\tilde{q}^o = \tilde{q}^o(\tilde{\varepsilon}_o/\ell) = q(x_o,p_o) = \frac{1}{\sqrt{2\pi}\,\tilde{\varepsilon}_o} \int_{-\ell/2}^{\ell/2} e^{-\xi^2/(2\tilde{\varepsilon}_o^{\,2})}\, d\xi \qquad 5.6$$

Since x_o is Gaussian, p_o maximizes the hit probability as well as it minimizes the error. Thus

$$q^o \leq \tilde{q}^o . \qquad 5.7$$

We can also define

$$\tilde{q}_o = \inf_x q(x,\tilde{p}_o) , \qquad 5.8$$

where $\tilde{p}_o$ is a variant of the algorithm p_o and will be described in Section 6.

As usual, we have

$$\tilde{q}_o \leq q_o \leq q^o \leq \tilde{q}^o , \qquad 5.9$$

but this time we have not been able to collapse this chain of inequalities. In fact we are only able to find a lower bound $z(\tilde{\epsilon}_o/\ell)$ such that

$$z \leq \tilde{q}_o \leq q_o \leq q^o \leq \tilde{q}^o . \qquad 5.10$$

Nevertheless, these bounds may well still be useful for first estimates since they provide a variation of no more than 70 percent. A tabulation of the lower bound z and the upper bound $\tilde{q}^o$ and their ratio is given in Table 1.

6. THE ALGORITHM $\tilde{p}_o$

Define the algorithm $\tilde{p}_o$ to be

$$\hat{x}(t+T,\tilde{p}_o) = y_o + \hat{x}(t+T,p_o). \qquad 6.1$$

Then

$$u(y \mid x,\tilde{p}_o) = u(y - y_o \mid x,p_o) . \qquad 6.2$$

The value y_o is the value that maximizes

$$q(y_o,x,p_o) = \int_{y=-\ell/2}^{y=\ell/2} du(y-y_o) \mid x,p_o) \qquad 6.3$$

From Appendix C we have

Table 1

$z(\varepsilon)$	ε	$\tilde{q}^o(\varepsilon)$	$\tilde{q}^o/z$
.00	∞	.000	1.382
.05	5.557	.072	1.434
.10	2.669	.149	1.486
.15	1.721	.229	1.524
.20	1.225	.317	1.585
.25	.935	.407	1.628
.30	.775	.481	1.605
.35	.622	.578	1.652
.40	.548	.639	1.597
.45	.461	.722	1.604
.50	.354	.843	1.685
.55	.335	.864	1.571
.60	.316	.886	1.477
.65	.296	.909	1.399
.70	.274	.932	1.332
.75	.250	.954	1.273
.80	.224	.975	1.218
.85	.194	.990	1.165
.90	.158	.998	1.109
.95	.112	1.000	1.053
1.00	.000	1.000	1.000

$$q(x,\tilde{p}_o) = \sup_{y_o} q(y_o,x,p_o) > z(\sigma/\ell) > z(\tilde{\varepsilon}_o/\ell) \qquad 6.4$$

where σ is the standard deviation around the mean. The last inequality follows since z is a monotonically decreasing function and since σ minimizes the r.m.s. error.

Since the middle inequality in Equation 6.4 holds for all x, we have

$$\tilde{q}_o > z(\tilde{\varepsilon}_o/\ell). \qquad 6.5$$

7. CONCLUSIONS

With respect to the r.m.s. criterion and the r.m.s. bound on an Nth derivative, the duel between a gunner and a target is a game with a saddle point which can be precisely defined and hence stable strategies exist for both players.

If hit probability is used as the criterion, we have been unable to define a saddle point precisely. However, we can find bounds that show that the difference between the performance for such a saddle point and the saddle point for the r.m.s. case may well be small enough to use the r.m.s. criterion as a good "first analysis."

APPENDIX A

First consider ε_o. We note that for each $\delta > 0$ there exists an x_δ such that

$$\inf_p \varepsilon(x_\delta,p) > \varepsilon_o - \delta$$

which follows from the definition of sup.

Thus

$$\varepsilon(x_\delta,p) > \varepsilon_o - \delta$$

for all p which follows from the definition of inf.

Likewise, there exists a p_δ such that

$$\varepsilon(x,p_\delta) < \varepsilon^o + \delta$$

for all x.

Thus

$$\varepsilon_o - \delta < f(x_\delta\, p_\delta) < \varepsilon^o + \delta$$

for all δ and thus

$$\varepsilon_o < \varepsilon^o.$$

APPENDIX B*

Statement of Problem.

Let k > 0 be such that the differential equation

$$x^{(2n)} = (-1)^n kx$$

$$x^{(m)}(T) = 0 \qquad (m=0,1,\dots,n-1)$$

$$x^{(m)}(0) = 0 \qquad (m=n,n+1,\dots,2n-1)$$

has a nontrivial solution. Prove that this solution is unique up to constant multiples.

Proof.

The proof consists in showing that for any two nontrivial solutions x and y the equality sign in Schwarz's inequality

$$\left(\int_0^T xy\, dt\right)^2 \leqq \int_0^T x^2 dt \int_0^T y^2 dt$$

holds, which occurs if and only if $y = cx$, c a constant.

To this end, we note first that

$$\int_0^T xy\, dt = \frac{1}{k} \int_0^T x^{(n)} y^{(n)} dt. \qquad \text{(B-1)}$$

This follows from

*The analysis in this appendix was provided by Mr. Walter O. Egerland of the Ballistic Modeling Division, Ballistic Research Laboratory.

$$\int_o^T xy\ dt = \frac{(-1)^n}{k}\int_o^T x^{(2n)}y\ dt$$

$$= \frac{(-1)^n}{k}\left\{ yx^{(2n-1)}\Big|_o^T - \int_o^T x^{(2n-1)}y'dt\right\}$$

$$= \frac{(-1)^{n+1}}{k}\int_o^T x^{(2n-1)}y'dt$$

$$= \frac{(-1)^{n+j}}{k}\int_o^T x^{(2n-j)}y^j dt,\ j=1,2,\ldots n \qquad \text{(B-2)}$$

Next, we write the identify

$$xy = x(0)y(0) + \int_o^t (x'y + xy')dt,$$

and find successively

$$xy = x(0)y(0) + \frac{(-1)^n}{k}\int_o^t (x'y^{(2n)} + y'x^{(2n)}dt$$

$$= x(0)y(0) + \frac{(-1)^n}{k}\left\{ y^{(2n-1)}x' + x^{(2n-1)}y' - \int_o^t (y^{(2n-1)}x'' + x^{(2n-1)}y'')dt\right\}$$

$$= \vdots$$

$$= x(0)y(0) + \frac{(-1)^n}{k}\left[\sum_{j=1}^{n-1} (-1)^{j+1}\left\{y^{(2n-j)}x^j + x^{(2n-j)}y^j\right\} + (-1)^{n+1}\, x^{(n)}\, y^{(n)}\right]$$

Integration over the interval [0,T], using (B-1) and (B-2), yields

$$\int_o^T xy\ dt = x(0)y(0)T - (2n-1)\int_o^T xy\ dt$$

or

$$\int_o^T xy\ dt = \frac{x(0)y(0)}{2n} T. \qquad \text{(B-3)}$$

In particular,

$$\int_o^T x^2 dt = \frac{x(0)^2}{2n} T$$

and

$$\int_o^T y^2 dt = \frac{y(0)^2}{2n} T.$$

Hence, equality in Schwarz's inequality holds, and the proof is complete.

APPENDIX C

Let $u(\xi)$ be any distribution function which we will take for convenience as having zero mean. Then

$$\sigma^2 = \int_{-\infty}^{\infty} \xi^2 \, du(\xi). \tag{C-1}$$

We want to relate σ with

$$q = \sup_{y} \int_{y-\ell/2}^{y+\ell/2} du \tag{C-2}$$

It is convenient to find a function such that

$$\sigma \geq r(q). \tag{C-3}$$

This function is monotonically decreasing with q and thus we can use it to define

$$q \geq z(\sigma). \tag{C-4}$$

implicitly.

We can write

$$\sigma^2 = \sum_{m=-\infty}^{\infty} \int_{m\ell/2}^{(m+1)\ell/2} \xi^2 \, du(\xi)$$

$$\sigma^2 \geq \sum_{m=0}^{\infty} \left(\frac{m\ell}{2}\right)^2 \int_{m\ell/2}^{(m+1)\ell/2} du + \sum_{m=-\infty}^{-1} \left(\frac{m+1}{2}\ell\right)^2 \int_{m\ell/2}^{(m+1)\ell/2} du$$

$$\sigma^2 \geq \sum_{m=0}^{\infty} \left(\frac{m\ell}{2}\right)^2 \mu_m + \sum_{m=-\infty}^{-1} \left[\frac{(m+1)\ell}{2}\right]^2 \mu_m \tag{C-5}$$

where

$$\mu_m = \int_{m\ell/2}^{(m+1)\ell/2} du \geq 0,$$

and
$$\mu_m + \mu_{m+1} \leqslant q$$

and
$$\sum_{m=-\infty}^{\infty} \mu_m = 1 .$$

If we write

$$\nu_m = \mu_m + \mu_{(-1-m)} ,$$

we have
$$\sigma^2 \geqslant \Sigma = \sum_{m=0}^{\infty} \frac{m\ell}{2}^{2} \nu_m , \qquad \text{(C-6)}$$

where
$$0 \leqslant \nu_o \leqslant q , \qquad \text{(C-7)}$$

$$0 \leqslant \nu_m + \nu_{m+1} \leqslant 2q , \qquad \text{(C-8)}$$

$$\sum_{m=0}^{\infty} \nu_m = 1. \qquad \text{(C-9)}$$

It is not hard to show that we have a lower bound for Σ defined in Equation C-6 if we let

$$\nu_m = q \quad \text{for } m=0,\ M-1, \qquad \text{(C-10)}$$

$$\nu_M = 1 - Mq, \qquad \text{(C-11)}$$

$$\nu_m = 0 \quad \text{for } m > M , \qquad \text{(C-12)}$$

where

$$M = \text{greatest integer } [1/q] . \qquad \text{(C-13)}$$

The proof of this goes as follows:

(i) If $\nu_o = q$, go to step (iv)

(ii) If $\nu_o < q$ and $\nu_o + \nu_1 = b \geqslant q$, put $\nu_o = q$ and $\nu_1 = b - q$. This will decrease Σ . Then go to step (iv).

(iii) If $\nu_o + \nu_1 = b < q$, put $\nu_o = b$, $\nu_1 = 0$ and reduce and other ν_m's to make $\nu_o = q$. This again will decrease Σ . Then go to step (iv).

(iv) Now work on ν_1 and ν_2 as we worked on ν_o and ν_1.

(v) Continue on with ν_2 and ν_3, etc.

We then have

$$\sigma^2/\ell^2 > \sum_{m=o}^{M-1} \frac{m^2 q}{4} + \frac{M^2}{4} [1 - Mq], \tag{C-14}$$

which we can write in closed form as

$$\sigma^2/\ell^2 = r(q) = \frac{(M-1)M(2M-1)}{24} q + \frac{M^2}{4} [1 - Mq]. \tag{C-15}$$

We now need only show that this function is monotonically decreasing for $0 < q \leq 1$, and we have implicitly defined $z(\sigma)$.

First let $q = \frac{1}{M}$. Then

$$r(q) = \frac{(M-1)(2M-1)}{24}$$

which increases as q decreases.

We can write Equation C-14 as

$$r(q) = \frac{M^2}{4} + qM\left[\frac{-4M^2 - 3M + 1}{24}\right] \tag{C-16}$$

For $\frac{1}{M+1} < q \leq \frac{1}{M}$, M is fixed in Equation C-16 and r(q) varies only as a constant times q. Since

$$-4M^2 - 3M + 1 < 0 \quad \text{for } q \leq 1,$$

we have that r(q) increases as q decreases from $\frac{1}{M}$ to $\frac{1}{M+1}$.

We should point out that $r(\cdot)$ is not actually achievable by a $u(\cdot)$. In particular, there is a jump of q at zero and a jump of $q/2$ at $\ell/2$ so the interval $(-\ell/4, 3\ell/4)$ would have measure $3q/2$. Thus z is only a lower bound.

STOCHASTIC MODELS, NON-LINEAR MODELS AND TIME-VARIABLE DETERMINISTIC MODELS OF COMBAT

Roger F. Willis
US Army TRASANA
White Sands Missile Range, NM 88002

ABSTRACT. This paper presents and solves several classes of deterministic and stochastic models of combat, including non-linear models and models with time-variable coefficients. The latter were developed to meet the need for simple, but more realistic, models of direct fire operations, in which weapon effectiveness and engagement opportunities vary with range. They can also be used for scenarios in which allocations of support resources are time-variable. The stochastic models can be completely solved in the sense that all required moments of the bivariate distribution of Red survivors and Blue survivors can be derived and plotted (as functions of time). From these moments the means and variances of selected measures of effectiveness can be derived.

1. INTRODUCTION. Combat models in general are either large complex simulations or very simple linear differential equation models with constant coefficients. This paper reports progress on model design efforts to bridge the gap between these two extremes. Three categories of models of intermediate complexity are covered: Solvable stochastic models, non-linear deterministic models and deterministic models with time-varying coefficients (called Bessel models of combat). Stochastic models are required because combat actually involves real uncertainties, the effects of which must be faced in comparisons of alternative systems or evaluations of alternative tactics. It has been found that non-linear models give a better match with the average results from stochastic simulations than linear models. This motivates our investigations of families of non-linear deterministic models. Models with time-varying coefficients are needed for more realistic representations of defender and attacker changes in detections and exposures as battles progress. For all three categories we investigate families of models with similar structure so we can take advantage of our knowledge about the uncertainties in the input data to guide us in switching from mathematically intractable models to nearby tractable models.

2. STOCHASTIC MODELS.

a. Laplace Transform Approach. The first type of stochastic model to be discussed involves combat between two homogeneous forces, the Red force and the Blue force. The basic assumptions are:

$V(a,b)\,\Delta t$ = probability of exactly one Red loss in time Δt, assuming that there are "a" Red and "b" Blue survivors.

$W(a,b)\,\Delta t$ = probability of exactly one Blue loss in time Δt, assuming that there are "a" Red and "b" Blue survivors.

The probability of more than one loss in time Δt, approaches zero as Δt approaches zero.

From these assumptions we derive the following set of differential equations, one for each pair (a,b) in the range $0 \le a \le M$, $0 \le b \le N$:

$$\frac{dp}{dt}(a,b) = V(a+1,b)p(a+1,b)+W(a,b+1)p(a,b+1) - p(a,b)[V(a,b)+W(a,b)]$$

the initial conditions are:

$$p(M,N,0) = 1$$

$$p(a,b,t) = 0 \quad \text{if } a>M \text{ or } b>N$$

$$p(a,b,0) = 0 \quad \text{if } a<M \text{ or } b<N$$

To solve this set of (M+1)(N+1) equations we take Laplace transforms of both sides and, using the initial conditions, derive the following recurrence relations between the successive Laplace transforms:

$$L(a,b) = \frac{W(a,b+1)L(a,b+1)+V(a+1,b)L(a+1,b)}{r+V(a,b)+W(a,b)}$$

By solving numerically for the rational functions that are the inverses of the transforms we derive the complete bivariate distribution of the Red survivors and Blue survivors, at any given time, as illustrated in Figure 1. This example is based on the following assumptions:

$$V(X,Y) = K\ Y = 0.268\ Y$$

$$W(X,Y) = J\ X = 0.100\ X$$

M = 3 (Blue weapons at time 0)

N = 9 (Red weapons at time 0)

b. Moment Generating Function Approach. We make the following assumptions:

$V\ B(t)\ \Delta t$ = probability of exactly one Red loss in time Δt

$W\ R(t)\ \Delta t$ = probability of exactly one Blue loss in time Δt

Probability of more than one loss in time Δt approaches zero as Δt approaches zero.

From these assumptions we can write down the partial differential equations satisfied by the moments of the bivariate distribution of Red survivors and Blue survivors at any given time (see page 118 of The Elements of Stochastic Processes, Bailey, N.T.J. (1964). New York: John Wiley). The equation is:

$$\frac{\partial M}{\partial t} = (e^{-X}-1)V\frac{\partial M}{\partial X} + (e^{-Y}-1)W\frac{\partial M}{\partial Y}$$

The bivariate moment generating function M is a function of X,Y and t (time) which, when expanded in a series of powers X^nY^m, has coefficients that are the moments (m_{ij}) of the distribution. The partial differential equation can be converted to a set of ordinary differential equations for the individual moments by expanding each of the functions $\frac{\partial M}{\partial t}$, $\frac{\partial M}{\partial X}$, $\frac{\partial M}{\partial Y}$, $e^{-X}-1$, $e^{-Y}-1$ into a series of powers of X^nY^m, performing the multiplications and additions required on the right-hand side of the equation, and finally equating coefficients of like powers of X^nY^m. For the first five moments, the following equations result:

$$\frac{dm_{10}}{dt} = -V\ m_{01}$$

$$\frac{dm_{01}}{dt} = -W\ m_{10}$$

$$\frac{dm_{11}}{dt} = -V\ m_{02} - W\ m_{20}$$

$$\frac{dm_{20}}{dt} = -2\ V\ m_{11} + V\ m_{01}$$

$$\frac{dm_{02}}{dt} = -2\ W\ m_{11} + W\ m_{10}$$

This set of five differential equations was solved analytically, yielding the results shown in Figure 2, in which the first five moments appear as explicit functions of the initial Red and Blue strengths and the kill rate coefficients V and W (which include assumptions about weapon characteristics, terrain, tactics, etc.). One can carry out sensitivity analyses directly with these functions since in many cases reasonable assumptions can be made about how the quantities V and W depend on key input assumptions, such as detection rates.

From these five moments means, variances and the covariance and correlation coefficient can be calculated, as follows:

$$\bar{R} = m_{10} \qquad \sigma_B^2 = m_{02} - m_{01}^2$$

$$\bar{B} = m_{01} \qquad \text{cov} = m_{11} - m_{10}m_{01}$$

$$\sigma_R^2 = m_{20} - m_{10}^2 \qquad \rho = \frac{\text{cov}}{\sigma_R\sigma_B}$$

The means and variances of measures of effectiveness, such as those listed below, can be calculated or estimated directly from these means, variances and the covariance. For example, the mean and variance of the difference in percent losses is:

$$\text{Mean} = \frac{m_{01}}{B_o} - \frac{m_{10}}{R_o}$$

$$\text{Variance} = \frac{m_{02} - m_{01}^2}{B_o^2} + \frac{m_{20} - m_{10}^2}{R_o^2} - \frac{2(m_{11} - m_{01}m_{10})}{B_oR_o}$$

A more complex stochastic model, between combatant forces X and Y, is based on the following assumptions:

PROBABILITY	CHANGE IN X	CHANGE IN Y
$\Delta t\ (a_1 X + p_1 Y)$	b_1	0
$\Delta t\ (a_2 X + p_2 Y)$	b_2	0
$\Delta t\ (a_3 X + p_3 Y)$	b_3	0
.	.	.
.	.	.
.	.	.
$\Delta t\ (a_N X + p_N Y)$	b_N	0
$\Delta t\ (d_1 X + r_1 Y)$	0	c_1
$\Delta t\ (d_2 X + r_2 Y)$	0	c_2
.	.	.
.	.	.
.	.	.
$\Delta t\ (d_M X + r_M Y)$	0	c_M

The partial differential equation for the moment generating function is:

$$\frac{\partial M}{\partial t} = \sum_{i=1}^{N} \left(e^{b_i X} - 1 \right) \left(a_i \frac{\partial M}{\partial X} + p_i \frac{\partial M}{\partial Y} \right) + \sum_{j=1}^{M} \left(e^{c_j Y} - 1 \right) \left(d_j \frac{\partial M}{\partial X} + r_j \frac{\partial M}{\partial Y} \right)$$

We have developed, and solved, a wide variety of stochastic models of combat, of which the above are typical examples.

3. NON-LINEAR DETERMINISTIC MODELS. It has been observed in the operations research literature several times that Lanchester models do not fit historical data and do not fit the average outcomes of complex stochastic simulations. One reason is that Lanchester models are usually assumed to be linear. It is possible to devise solvable non-linear models of combat that do fit historical data or simulation outputs. One way to approach this fitting process is to plot Red percent survivors versus Blue percent survivors (for example, using the average results from a number of replications with a stochastic simulation) and superimpose the solutions from a family of non-linear models to determine which members of the family come closest to the simulation results. One set of non-linear models is shown in Figure 3. As usual in such models R(t) is the Red strength surviving at time t, B(t) is the Blue strength surviving at time t and the indicated differentiations are with respect to time. In order to discuss the nature of the solutions we define R_0 and B_0 as the initial strengths, at time zero. In many cases the form of the solutions will depend on the sign of the constant A, where

$$A = \frac{2J}{K} R_0 - B_0^2$$

In a typical case B(t) is the square root of A times the tangent of a quadratic function of time, if A is positive, and the square root of -A times the hyperbolic tangent of a quadratic function of time, if A is negative. Five more solvable non-linear models, with constant coefficients, are presented in Figure 4. Figure 5 contains four solvable models that are non-linear with variable coefficients. These models also have solutions that are tangents or hyperbolic tangents with polynomial arguments. Such models could be used to represent time-correlated movement of assaulting forces, taking into account the density of targets in the force.

4. DETERMINISTIC MODELS WITH TIME-VARYING COEFFICIENTS. The first set of models with time-variable coefficients are listed in Figure 6. These and the other Bessel models presented in Figure 7 can all be solved analytically, yielding solutions that are either Bessel functions of fractional order or hyperbolic sines and cosines with polynomial arguments.

FIGURE 1

Probability of "a" Red Survivors and "b" Blue Survivors

T = .5 min.

b \ a	9	8	7	6	5	4	3	2	1	0
3	.427	.176	.036	.005	.001	.000	.000	.000	.000	.000
2	.205	.066	.011	.001	.000	.000	.000	.000	.000	.000
1	.050	.012	.001	.000	.000	.000	.000	.000	.000	.000
0	.008	.001	.000	.000	.000	.000	.000	.000	.000	0.0

T = 2 min.

b \ a	9	8	7	6	5	4	3	2	1	0
3	.033	.059	.053	.031	.014	.005	.002	.000	.000	.000
2	.079	.108	.074	.034	.012	.003	.001	.000	.000	.000
1	.094	.094	.047	.016	.004	.001	.000	.000	.000	.000
0	.074	.048	.017	.004	.001	.000	.000	.000	.000	0.0

T = 6 min.

b \ a	9	8	7	6	5	4	3	2	1	0
3	.000	.000	.000	.002	.003	.004	.004	.004	.003	.002
2	.000	.002	.006	.010	.013	.013	.010	.007	.004	.002
1	.003	.012	.021	.025	.023	.016	.009	.004	.001	.000
0	.015	.031	.037	.030	.019	.009	.003	.001	.000	0.0

FIGURE 2

$$\boxed{m_{10}} = R_o \cosh(t\sqrt{VW}) - B_o\sqrt{\frac{V}{W}} \sinh(t\sqrt{VW})$$

$$\boxed{m_{01}} = B_o \cosh(t\sqrt{VW}) - R_o\sqrt{\frac{W}{V}} \sinh(t\sqrt{VW})$$

$$\boxed{m_{11}} = \frac{R_o + B_o}{3}\left[\cosh(t\sqrt{VW}) - \cosh(2t\sqrt{VW})\right]$$

$$+ R_oB_o \cosh(2t\sqrt{VW}) - \frac{1}{3}\left(-\sqrt{\frac{V}{W}}\,B_o + \sqrt{\frac{W}{V}}\,R_o\right) \sinh(t\sqrt{VW})$$

$$+ \frac{1}{6}\left\{\sqrt{\frac{V}{W}}\left(B_o - 3B_o^2\right) + \sqrt{\frac{W}{V}}\left(R_o - 3R_o^2\right)\right\} \sinh\left(2t\sqrt{VW}\right)$$

$$\boxed{m_{20}} = \frac{1}{3}\left(\frac{2V}{W}\,B_o - R_o\right) \cosh(t\sqrt{VW})$$

$$+ \frac{1}{3}\sqrt{\frac{V}{W}}\left(B_o - 2R_o\right) \sinh(t\sqrt{VW}) + \sqrt{\frac{V}{W}}\left(\frac{R_o + B_o}{3} - R_oB_o\right) \sinh(2t\sqrt{VW})$$

$$- \frac{1}{6}\left\{\frac{V}{W}\left(B_o - 3B_o^2\right) + \left(R_o - 3R_o^2\right)\right\} \cosh(2t\sqrt{VW})$$

$$+ \frac{1}{2}\left\{R_o^2 + R_o - \frac{V}{W}\left(B_o^2 + B_o\right)\right\}$$

FIGURE 2 (CONTINUED)

$$\boxed{m_{02}} = \frac{1}{3}\left(\frac{2\,W}{V} R_o - B_o \right) \cosh(t\sqrt{WV})$$

$$+ \frac{1}{3} \sqrt{\frac{W}{V}} \left(R_o - 2B_o \right) \sinh(t\sqrt{WV}) + \sqrt{\frac{W}{V}} \left(\frac{R_o + B_o}{3} - R_o B_o \right) \sinh(2t\sqrt{VW})$$

$$- \frac{1}{6} \left\{ \frac{W}{V} \left(R_o - 3R_o^2 \right) + \left(B_o - 3B_o^2 \right) \right\} \cosh(2t\sqrt{VW})$$

$$+ \frac{1}{2} \left\{ B_o^2 + B_o - \frac{W}{V} \left(R_o^2 + R_o \right) \right\}$$

FIGURE 3

NON-LINEAR; CONSTANT COEFFICIENTS

$\dot{R}=-KBR$ $\dot{B}=-JR$	$\dot{R}=-KBR$ $\dot{B}=-JRB$	$\dot{R}=-KBR$ $\dot{B}=-JRB^{2}$
$\dot{R}=-KRB^{2}$ $\dot{B}=-JR$	$\dot{R}=-KRB^{2}$ $\dot{B}=-JRB$	$\dot{R}=-KBR^{-1/2}$ $\dot{B}=-JBR^{3/2}$

FIGURE 4

NON-LINEAR; CONSTANT COEFFICIENTS

$$\dot{R}=-KB^{1.5}R^{0.5}$$

$$\dot{B}=-JR^{0.5}B^{1.5}$$

$$\dot{R}=-KBR+aR$$

$$\dot{B}=-JRB+bB$$

$$\dot{R}=-KBR-cB$$

$$\dot{B}=-JRB-dR$$

$$\dot{R}=-fB^2R-cBR+aR$$

$$\dot{B}=-hR^2B-gRB+bB$$

$$\dot{R}=(f+gB)(q+aR)$$

$$\dot{B}=(h+kR)(p+bB)$$

FIGURE 5

NON-LINEAR; VARIABLE COEFFICIENTS

$\dot{R}=-KtBR$ $\dot{B}=-JtR$	$\dot{R}=-Kt^{N}BR$ $\dot{B}=-Jt^{N}R$
$\dot{R}=-K(a+bt)BR$ $\dot{B}=-J(a+bt)R$	$\dot{R}=-K(a+bt+ct^{2})BR$ $\dot{B}=-J(a+bt+ct^{2})R$

FIGURE 6

VARIABLE COEFFICIENT MODELS

$\dot{R} = -AtB(t)$ $\dot{B} = -CtR(t)$	$\dot{R} = -(At+C)B(t)$ $\dot{B} = -(aAt+aC)R(t)$
$\dot{R} = -AB$ $\dot{B} = -CtR$	$\dot{R} = -AtB$ $\dot{B} = -CR$
$\dot{R} = -AB$ $\dot{B} = -(Ct+b)R$	$\dot{R} = -AB$ $\dot{B} = -Ct^2 R$

FIGURE 7

VARIABLE COEFFICIENT MODELS

$\dot{R} = -At^{N}B$ $\dot{B} = -Ct^{N}R$	$\dot{R} = -At^{N}B$ $\dot{B} = -Ct^{P}R$
$\dot{R} = -Ae^{NT}B$ $\dot{B} = -Ce^{NT}R$	$\dot{R} = -Ae^{NT}B$ $\dot{B} = -Ce^{PT}R$

A SOLUTION FOR NON-COOPERATIVE GAMES

Prakash P. Shenoy
Mathematics Research Center
University of Wisconsin at Madison
Madison, WI 53706

ABSTRACT. In this paper we study solutions of strict non-cooperative games that are played just once. The players are not allowed to communicate with each other. The main ingredient of our theory is the concept of rationalizing a set of strategies for each player of a game. We state an axiom based on this concept that every solution of a non-cooperative game is required to satisfy. Strong Nash solvability is shown to be a sufficient condition for the rationalizing set to exist, but it is not necessary. Also, Nash solvability is neither necessary nor sufficient for the existence of the rationalizing set of a game. For a game with no solution (in our sense), a player is assumed to recourse to a "standard of behavior". Some standards of behavior are examined and discussed.

I. INTRODUCTION. In this paper, we study solutions of non-cooperative games. In a non-cooperative game, absolutely no preplay communication is allowed between the players. The theory of non-cooperative games, in contrast with cooperative games, is based on the absence of coalitions in that it is assumed that each participant acts independently without collaboration or communication with any of the others. Since in repeated plays of a game it is possible for players to "communicate" or signal via their choice patterns on previous plays[†] we shall avoid this feature of a non-cooperative game by only considering games that are played just once. Our objective is to study strict non-cooperative games and although this may be a severe restriction on the class of realistic games, like Luce and Raiffa [6, pp. 105], we feel that

> "...the realistic cases actually lie in the hiatus between strict non-cooperation and full cooperation but that one should first attack these polar extremes."

Besides, in many of the games that arise in the military and political contexts, the players often have a single-play orientation.

Except for this difference, we make the usual assumptions of rationality and complete information, i.e., all players are "rational"[‡] and each player has complete information of this fact and of his own and other players' utility function.

[†] See Luce and Raiffa [6, pp. 97-102] for a discussion of the temporal repetition of the prisoner's dilemma.

[‡] Here we mean in the usual von Neumann and Morgenstern sense. Later in Section III, we will look at this assumption more critically and study its implications.

Sponsored by the United States Army under Contract No. DAAG29-75-C-0024 and by the National Science Foundation under Grant No. MCS75-17385 A01.

II. *FORMAL DEFINITIONS AND TERMINOLOGY*. In this section we will define the basic concepts in the non-cooperative theory. The non-cooperative idea will be implicit, rather than explicit, below.

An *n-person game* is a set of n *players* denoted by $N = \{1,\dots,n\}$, each with an associated finite set of *pure strategies*; and corresponding to each player, i, a *von Neumann-Morgenstern utility function* u_i, which maps the set of all n-tuples of pure strategies into real numbers. By the term n-tuple, we mean a set of n items with each item associated with a different player. A *mixed strategy* of player i will be a probability distribution on his set of pure strategies. We write $s^i = \sum_\alpha c_{i\alpha}\pi_{i\alpha}$ with $c_{i\alpha} \geq 0$ and $\sum_\alpha c_{i\alpha} = 1$ to to represent such a mixed strategy, where the $\pi_{i\alpha}$'s are the pure strategies of player i. The von Neumann-Morgenstern utility function u_i used in the definition of a finite game above has a unique extension to the n-tuples of mixed strategies which is linear in the mixed strategy of each player (n-linear). This extension we will also denote by u_i, writing $u_i(s^1,s^2,\dots,s^n)$. I.e.,

$$u_i(s^1,s^2,\dots,s^n) = \sum_{\alpha_1} \dots \sum_{\alpha_n} c_{1\alpha_1}\dots c_{n\alpha_n} u_i(\pi_{1\alpha_1},\dots,\pi_{n\alpha_n}) .$$

We shall use the symbols i,j,k for players and α,β,γ to indicate various pure strategies of a player. The symbols s^i,t^i,r^i will indicate mixed strategies; $\pi_{i\alpha}$ will denote the i^{th} player's α^{th} pure strategy, etc. We shall write $\bar{s},\bar{t}$ to denote an n-tuple of mixed strategies. For convenience we shall use the substitution notation $(\bar{s};t_i)$ to denote $(s^1,\dots,s^{i-1},t^i,s^{i+1},\dots,s^n)$ where $\bar{s} = (s^1,\dots,s^n)$.

An n-tuple $\bar{s}$ is a *Nash equilibrium point* if and only if for every i

$$u_i(\bar{s}) = \max_{\text{all } t^i\text{'s}} [u_i(\bar{s};t^i)] .$$

Thus an equilibrium point is an n-tuple $\bar{s}$ such that each player's mixed strategy maximizes his payoff if the strategies of the others are held fixed. In an extremely elegant proof, Nash [8] has shown that every non-cooperative game with finite sets of pure strategies has an equilibrium point. A strategy s^i is player i's *equilibrium strategy* if the n-tuple $(\bar{t};s^i)$ is an equilibrium point for some n-tuple $\bar{t}$.

A strategy r^i is player i's *maximin* strategy if and only if for all n-tuples $\bar{s}$,

$$u_i(\bar{s};r^i) \geq \max_{\text{all } s^i\text{'s}} \ \min_{\text{all } s^1,\dots,s^{i-1},s^{i+1},\dots,s^n} [u_i(s^1,\dots,s^n)] .$$

The quantity on the right side of the above inequality is called player i's *maximin value* and denoted by v_i^m.

For 2-person games only, a strategy t^i is player i's *minimax* strategy if and only if for all player j's strategies, s^j, $j \neq i$

$$u_j(t^i,s^j) \leq \min_{\text{all } s^i\text{'s}} \max_{\text{all } s^j\text{'s}} [u_j(s^i,s^j)] .$$

We say that a mixed strategy s^i *uses* a pure strategy $\pi_{i\alpha}$ if $s^i = \sum_\beta c_{i\beta}\pi_{i\beta}$ and $c_{i\alpha} > 0$. If $\bar{s} = (s^1,\ldots,s^n)$ and s^i uses $\pi_{i\alpha}$, we also say that $\bar{s}$ uses $\pi_{i\alpha}$. Let s^i and r^i be two distinct mixed strategies for player i. We say s^i *strongly dominates* r^i if $u_i(\bar{t};s^i) > u_i(\bar{t};r^i)$ for every $\bar{t}$. This amounts to saying that s^i gives player i a higher payoff than r^i no matter what the strategies of the other players are. To see whether a strategy s^i strongly dominates r^i, it suffices to consider only pure strategies for the other players because of the n-linearity of u_i. Also, we say s^i *weakly dominates* r^i if $u_i(\bar{t};s^i) \geq u_i(\bar{t};r^i)$ for all $\bar{t}$ and strict inequality holds for at least one $\bar{t}$.

Based on the concept of an equilibrium point, Nash defined several "solutions" of non-cooperative games. A game is said to be *Nash solvable* if its set S of equilibrium points satisfies the condition

$$(\bar{t};r^i) \in S \text{ and } \bar{s} \in S \Rightarrow (\bar{s};r^i) \in S . \tag{2.1}$$

This is called the *interchangeability* condition. The *Nash solution* of a Nash solvable game is its set S of equilibrium points. A game is *strongly Nash solvable* if it has a Nash solution, S, such that for all i's

$$\bar{s} \in S \text{ and } u_i(\bar{s};r^i) = u_i(\bar{s}) \Rightarrow (\bar{s};r^i) \in S$$

and then S is called a *strong Nash solution*. If S is a subset of the set of equilibrium points of a game and satisfies condition (2.1); and if S is maximal relative to this property, then we call S a *Nash subsolution*. Let S be the set of all equilibrium points of a game. Define

$$v_i^+ = \max_{\bar{s}\in S} [u_i(\bar{s})],, \quad v_i^- = \min_{\bar{s}\in S} [u_i(\bar{s})] .$$

If $v_i^+ = v_i^-$, we write $\mathbf{v}_i = v_i^+ = v_i^-$. v_i^+ is called the *Nash upper value* to player i of the game; v_i^- the *Nash lower value*; and $\mathbf{v}_i$ the *Nash value*, if it exists.

Note that a non-cooperative game does not always have a Nash solution, but when it does, the Nash solution is unique. Strong Nash solutions are Nash solutions with special properties. Nash subsolutions always exist and have many of the properties of Nash solutions, but lack uniqueness. A Nash subsolution, when unique, is a Nash solution.

Apart from these "solutions", Luce and Raiffa [6, Ch. 5] have defined "solution in the strict sense", "solution in the weak sense" and "solution in the complete weak sense". For reasons of space, we do not repeat these definitions here.

A natural question that arises is: In what sense are these concepts, solutions of non-cooperative games? I.e., what constitutes a solution of a non-cooperative game? These questions are discussed in the subsequent sections.

III. SOLUTIONS OF NON-COOPERATIVE GAMES. What do we mean by a solution of a non-cooperative game? Let Γ be a n-person non-cooperative game. Consider player i's position in this game. He is informed about the pure strategy sets of all the players. He is also aware of the von Neumann-Morgenstern utilities of all players associated with every possible n-tuple of pure strategies. The only other information he has about the other players is that they are rational players. The game is to be played just once. Given all these facts, which strategy should he play in order to maximize his utility? In this situation, if a logical analysis of the problem requires player i to play a particular strategy or a strategy from a particular set of strategies, such a course of action can be called a solution for player i. On the other hand, a logical analysis of the situation under the given set of information may not lead to any particular conclusion, in which case we can say that for the given game, there is no solution for player i. In the latter case, assuming that not playing the game is not one of the options that player i has, player i is still faced with the question of having to pick a strategy. We will assume that in this case player i recourses to a "standard of behavior" (as distinct from a solution) to pick a strategy from the set of all his strategies. Which standard of behavior player i should opt for is then clearly a meta-game theoretical question and beyond the scope of game theory.

We will now attempt to define a solution for a non-cooperative game (if one exists). Consider again player i's situation in a game. If he had prior information about the strategies that his opponents would employ, his problem of selecting a strategy would simplify to finding the strategy which would maximize his utility subject to the restriction that each of his opponents play a fixed strategy which is known to player i. However, player i has no such prior information. The only clue he has about the actions of the other players is the fact that they are rational players. What does the assumption of rationality imply about players' behavior?

One implication is that if for some player k, his pure strategy $\pi_{k\alpha}$ is strongly dominated by another pure strategy $\pi_{k\beta}$, then player k has never any incentive to play a mixed strategy that uses the pure strategy $\pi_{k\alpha}$. This is because, no matter what strategies the other players play, player k can do better by playing instead the mixed strategy obtained by substituting $\pi_{k\beta}$ in place of $\pi_{k\alpha}$. Thus a given game can be reduced by the elimination of all strongly dominated pure strategies of all the players. The reduced game is again examined for strongly dominated pure strategies and the process continued until no player has a strongly dominated pure strategy.

What else can we deduce from the assumption of rationality? We examine this first for 2-person games. If player i plays a mixed strategy s^{*i}, then the best reply for the other player, j, is to play any strategy from the set

$$M_j(s^{*i}) = \{s^{*j} : u_j(s^{*j},s^{*i}) = \max_{s^j} u_j(s^j,s^{*i})\} . \tag{3.1}$$

Similarly, if player j plays a mixed strategy s^{*j}, the best reply for player i is to play a strategy from the set $M_i(s^{*j})$ defined as in (3.1). Suppose, on the basis of the assumption of rationality, we can rationalize a unique strategy s^{*i} for player i. I.e., we suppose that, since player i is a rational player, he is expected to play a particular strategy s^{*i} (and no other). Then, since player j is also a rational player, we can rationalize the set of strategies $M_j(s^{*i})$ for player j. I.e., player j can be expected to play any strategy from the set $M_j(s^{*i})$. Then, if our original assumption of rationalizing s^{*i} for player i is to be valid, we must have

$$\{s^{*i}\} = M_i(s^j) \ \forall \ s^j \in M_j(s^{*i}) .$$

In general, we may be able to rationalize a (unique) set of strategies for each player. We make the following formal definition for a 2-person game. A non-empty set of strategies X^i can be <u>rationalized</u> for player i if and only if it is the unique set satisfying the following two conditions:

$$\exists \ X^j \text{ such that } X^j = M_j(s^i) \ \forall \ s^i \in X^i \tag{3.2}$$

$$X^i = M_i(s^j) \ \forall \ s^j \in X^j . \tag{3.3}$$

The following proposition is an obvious consequence of the above definition.

<u>Proposition 3.1</u>. If X^i can be rationalized for player i, then X^j given by (3.2) can be rationalized for player j.

<u>Proof:</u> Since conditions (3.2) and (3.3) are valid, we only need to show that X^j is a unique set satisfying these conditions. This follows from the fact that X^i is a unique set satisfying these conditions.

Q.E.D.

The concept of rationalizing a set of strategies for each player in a 2-person game can easily be generalized to a n-person game. Let

$$M_i(s^1,\dots,s^{i-1},s^{i+1},\dots,s^n) = \{t^i : u_i(\bar{s};t^i) = \max_{\text{all } r^i\text{'s}} [u_i(\bar{s};r^i)]\}$$

$$\text{where } \bar{s} = (s^1,\dots,s^n) \ .$$

Let Γ be an n-person game. Let $X = (X^1,\dots,X^n)$ be an n-tuple of nonempty sets of strategies. We say X can be *rationalized* for Γ (or X^i can be *rationalized for player* i, $i = 1,\dots,n$) if X is the unique n-tuple satisfying for all $i \in N$

$$X^i = M_i(s^1,\dots,s^{i-1},s^{i+1},\dots,s^n) \ \forall \ (s^1,\dots,s^{i-1},s^{i+1},\dots,s^n)$$

$$\in X^1 \times \dots \times X^{i-1} \times X^{i+1} \times \dots \times X^n \ .$$

Thus we see that the concept of rationalizing an n-tuple of sets of strategies for a game is a minimal condition that every solution of a non-cooperative game should satisfy, i.e., it is a "necessary" condition. We will now attempt to show that it is, in a sense, a "sufficient" condition as well.

Consider a 2-person game such that we can rationalize X^i for player i and X^j for player j. Player i's situation can be summarized as in Table 1. Hence player i has a reasonable justification for playing a strategy from the set X^i. Also if player j anticipates this action of player i, his subsequent action merely reinforces player i's *choice of picking a strategy from* X^i. A similar argument can be made for player i if the game has n players.

If player i picks a strategy from the set	and	Assuming that player j picks a strategy from the set	then	The utility payoff to player i is
X^i		X^j		the best that player i can hope for
		$(X^j)^c$		indeterminate
$(X^i)^c$		X^j		worse off than if player i had played a strategy from X^i
		$(X^j)^c$		indeterminate

Table 1

We have stated two implications of rationality. We can consider these as axioms that a solution of a non-cooperative game should always satisfy (if one exists). For example,

Axiom 0: A non-cooperative game may or may not have a solution.

Axiom 1: If a non-cooperative game has a solution and $\bar{s}$ is an n-tuple of strategies in the solution, then $\bar{s}$ does not use any strongly dominated strategy.

Axiom 2: If a non-cooperative game has a solution, then it should be rationalizable for the game.

It is clear from the definitions that a rationalizable set cannot contain a strategy that uses a strongly dominated strategy. Hence Axiom 2 implies Axiom 1. In the next section, we examine Nash's various solutions and see how they relate to our axioms.

IV. THE ROLE OF EQUILIBRIUM POINTS IN SOLUTIONS OF NON-COOPERATIVE GAMES.

The concept of a Nash equilibrium point is the basic ingredient of Nash's theory of non-cooperative games. We will show that it also plays an important role in our theory.

Proposition 4.1. Let X be rationalizable for Γ. Then $\bar{s} \in X \Rightarrow \bar{s}$ is a Nash equilibrium point.

The proof follows from the definition of a rationalizable set for Γ. We now examine Nash's theory of non-cooperative games and see how they relate to our axioms.

Theorem 4.2: Let Γ be a strongly Nash solvable game. Then the strong Nash solution S is rationalizable for Γ.

Proof: Let $X^i = \{r^i : (\bar{s};r^i) \in S$ for some $\bar{s}\}$. Clearly

$$X^i \subset M_i(s^1,\ldots,s^{i-1},s^{i+1},\ldots,s^n) \ \forall\ (s^1,\ldots,s^{i-1},s^{i+1},\ldots,s^n)$$

$$\in X^1 \times \ldots \times X^{i-1} \times X^{i+1} \times \ldots \times X^n .$$

Since Γ is strongly Nash solvable,

$$\bar{s} \in S\ ,\ u_i(\bar{s};r^i) = u_i(\bar{s}) \Rightarrow (\bar{s};r^i) \in S\ .$$

So we have

$$X^i \supset M_i(s^1,\ldots,s^{i-1},s^{i+1},\ldots,s^n) \ \forall\ (s^1,\ldots,s^{i-1},s^{i+1},\ldots,s^n)$$

$$\in X^1 \times \ldots \times X^{i-1} \times X^{i+1} \times \ldots \times X^n .$$

Hence

$$X^1 = M_i(s^1,\dots,s^{i-1},s^{i+1},\dots,s^n) \ \forall \ (s^1,\dots,s^{i-1},s^{i+1},\dots,s^n)$$

$$\in X^1 \times \dots \times X^{i-1} \times X^{i+1} \times \dots \times X^n .$$

Hence $X = (X^1,\dots.X^n)$ is rationalizable for Γ. But $X = \mathcal{S}$. Hence $\mathcal{S}$ is rationalizable for Γ.

Q.E.D.

Theorem 4.2 states that strong Nash solvability is a sufficient condition for the existence of a rationalizable set and that the rationalizable set coincides with the strong Nash solution. However, the surprising result is that strong Nash solvability is not a necessary condition for the existence of a rationalizable set. The following example illustrates this fact.

Example 4.1: Consider the 2-person game represented by the matrix given below

$$\begin{array}{cc|cc|}
 & & \multicolumn{2}{c}{2} \\
 & & \beta_1 & \beta_2 \\
\hline
\text{1:} & \alpha_1 & (1,3) & (1,3) \\
 & \alpha_2 & (0,0) & (2,2) \\
\hline
\end{array}$$

The equilibrium points of this game are (α_1,β_1) and (α_2,β_2). These are not interchangeable, hence the game is not even Nash solvable. However, it can easily be shown that $\{(\alpha_2,\beta_2)\}$ is rationalizable for the game. □

Since the game in Example 4.1 is not Nash solvable, Nash solvability is not a necessary condition for the existence of the rationalizable set. Moreover, Nash solvability is not a sufficient condition for the existence of a rationalizable set. This is shown in the next example.

Example 4.2: Consider the 2-person game represented by the matrix given below

$$\begin{array}{cc|cc|}
 & & \multicolumn{2}{c}{2} \\
 & & \beta_1 & \beta_2 \\
\hline
\text{1:} & \alpha_1 & (5,-3) & (-4,5) \\
 & \alpha_2 & (-5,5) & (3,-4) \\
\hline
\end{array}$$

This game has a unique equilibrium point $(\frac{9}{16}\alpha_1 + \frac{7}{16}\alpha_2, \frac{7}{17}\beta_1 + \frac{10}{17}\beta_2)$. Thus the game is Nash solvable. The Nash value of the game to player 1 is -5/17 and to player 2 is 1/2. It can easily be shown that the rationalizable set does not exist for this game. Hence from our point of view, the game has no solution. To see why Nash's solution is not really a solution of this game, consider player 2's position. If he plays his equilibrium strategy, the maximum he can get is his Nash value, 1/2, provided player 1 also plays his equilibrium strategy. However, player 2 can guarantee his Nash value irrespective of player 1's actions by simply playing the maximin strategy $(\frac{1}{2}\beta_1 + \frac{1}{2}\beta_2)$. Moreover, if player 2 plays his equilibrium strategy and player 1 plays his maximin strategy $(\frac{8}{17}\alpha_1 + \frac{9}{17}\alpha_2)$ (to guarantee his Nash value, -5/17), player 2 actually gets 107/289 which is less than his Nash value!

On the subject of rational behavior, von Neumann and Morgenstern [9] write:

> "... the rules of rational behavior must provide definitely for the possibility of irrational conduct on the part of others... . If that should turn out to be advantageous for them - and quite particularly, disadvantageous to the conformists then the above "solution" would seem very questionable".

Hence it is not clear why player 2 should play his equilibrium strategy. □

Next, we study the implications of our axioms when applied to the special and well known case of 2-person zero-sum games. We say a 2-person zero-sum game has a saddle point if it has an equilibrium point in pure strategies. I.e. if $\exists\ \pi_{i\alpha}, \pi_{j\beta}$ such that $(\pi_{i\alpha}, \pi_{j\beta})$ is an equilibrium point.

Proposition 4.3. Let Γ be a 2-person zero-sum game. *The game has a rationalizable set only if* Γ has a saddle point.

Proof: If Γ has a saddle point such that it is a strong Nash solution, then by Theorem 4.2 it is rationalizable for Γ. If Γ has no saddle point, then there exists a unique Nash equilibrium in mixed strategies. If player i plays his equilibrium strategy, then player j can play any pure strategy used in his equilibrium strategy and still get his Nash value of the game and vice-versa. Hence $\exists$ no rationalizable set for the game.

Q.E.D.

Thus, as per our theory, a 2-person zero-sum game with no saddle point has no solution. This is in sharp contrast with the universally accepted theory of von Neumann and Morgenstern [9] that the equilibrium point always constitutes a solution of a 2-person zero-sum game. Although we agree that there are many other reasons why a player may want to play the equilibrium strategy[†], we feel that it is not necessarily a consequence of the assumption of rationality of the players.

Since the rationalizable set does not always exist, we cannot have a general existence result. However, this should not be interpreted negatively. I.e. a lack of a general existence result is not a "defect" in our theory. It is merely

[†] Some of these reasons are discussed in Section V of this paper.

an outcome of the "lack of information" that a player has in playing certain non-cooperative games. I.e. some games, those for which a rationalizable set does not exist, do not give sufficient insight into the behavior of players assuming only rationality. We do not believe that the conditions imposed by Axiom 2 are too strong and must therefore be modified to admit existence for all games. We feel that Axiom 2 is a minimal condition that every solution should satisfy. For a game that has no solution (in our sense), a player can recourse to a "standard of behavior". These are discussed in the next section.

V. SOME STANDARDS OF BEHAVIOR. Let Γ be a game that has no rationalizable set. Consider the position of a player, i. He has to pick a strategy to maximize his utility. His job is complicated by the fact that since the rationalizable set does not exist, he has no inkling of the strategies that the other players are going to pick. Some of the possible actions that he can take are as follows.

Undominated Strategies.

The fact that the game has no rationalizable set does not exclude the fact that some player(s) may have strongly dominated pure strategies. If this is the case, it is safe to assume that a player will never use a strongly dominated pure strategy in any mixed strategy and thus the game can be reduced by the elimination of all strongly dominated pure strategies. The reduced game is again examined for strongly dominated pure strategies and the process continued until no player has a strongly dominated pure strategy. At the end of this reduction process, since the game has no rationalizable set, there will be at least 2 players each of whom will have at least 2 pure strategies.

Let Γ be a game with no rationalizable set and no strongly dominated pure strategy. Suppose some player, j, has a weakly dominated pure strategy. Since player j can do as well (if not better) by substituting the weakly dominated pure strategy by the dominating pure strategy in any mixed strategy that uses such a weakly dominated strategy, it is conceivable that he will never use his weakly dominated pure strategy in any mixed strategy. Thus the game can be reduced by the elimination of all weakly dominated strategies. By the same reasoning, the reduced game is again examined for weakly dominated strategies and the process continued until no player has a weakly dominated strategy.

Maximin Strategies.

In a finite game, maximin strategies always exist for all players. Let Γ be a game for which no rationalizable set exists. Also suppose that no player has a dominated pure strategy. For such games, since a player has no idea of the strategies that the other players will play, he may decide to protect himself as much as possible by playing the maximin strategy. Thus by playing a maximin strategy, a player, i, is assured of getting at least his maximin value v_i^m irrespective of the actions of the other players.

For 2-person zero-sum games, a player's maximin strategy is also his minimax strategy since

$$\begin{aligned}\max_{s^i}\min_{s^j}[u_i(s^i,s^j)] &= \max_{s^i}\min_{s^j}[-u_j(s^i,s^j)]\\ &= \max_{s^i}\{-\max_{s^j}[u_j(s^i,s^j)]\}\\ &= -\{\min_{s^i}\max_{s^j}[u_j(s^i,s^j]\}\ .\end{aligned}$$

Also since for all 2-person zero-sum games,

$$v_i^m = -v_j^m$$

a player's maximin strategy is also his equilibrium strategy. Thus, in a 2-person zero-sum game, there is a strong motivation for a player to play his maximin (which is also his minimax and equilibrium) strategy. However, as mentioned before, we are not willing to subscribe to the theory that this constitutes a solution of the game.

In general, for 2-person non-zero-sum games, maximin strategies are distinct from equilibrium strategies and often the maximin value of a player is equal to the Nash value (when it exists). In such cases we feel that it is better in some respects for a player to play his maximin strategy instead of his equilibrium strategy.

Minimax Strategies in 2-Person Games.

For 2-person non-zero-sum games, minimax strategies are usually distinct from maximin strategies. However they often coincide with equilibrium strategies. Since in a non-zero-sum game, the utility of an outcome for a player has no relation to the utility of the same outcome to his opponent, we cannot see any motivation for a rational player to play his minimax strategy (on its' merits alone).

Equilibrium Strategies.

Since equilibrium points always exist, every player i has a nonempty set S_i of equilibrium strategies. The concept of an equilibrium strategy alone is not strong enough to qualify even as a standard of behavior. E.g., for games that are not Nash solvable, it makes no sense for a player to play an equilibrium strategy because the resulting outcome may not be an equilibrium point. For games that are Nash solvable (but not strongly Nash solvable) equilibrium strategies may qualify as a standard of behavior.

We end this section by discussing a 2-person non-zero-sum game in detail.

			2			
			equilibrium	equilibrium	equilibrium and minimax	maximin
			β_1	β_2	$3/8\ \beta_1 + 5/8\ \beta_2$	$5/8\ \beta_1 + 3/8\ \beta_2$
1	equilibrium	α_1	(1,2)	(-1,-4)	(-1/4,-7/4)	(1/4,-1/4)
	equilibrium	α_2	(-4,-1)	(2,1)	(-1/4,1/4)	(-7/4,-1/4)
	equilibrium and minimax	$1/4\ \alpha_1 + 3/4\ \alpha_2$	(-11/4,-1/4)	(5/4,-1/4)	(-1/4,-1/4)	(-5/4,-1/4)
	maximin	$3/4\ \alpha_1 + 1/4\ \alpha_2$	(-1/4,5/4)	(-1/4,-11/4)	(-1/4,-5/4)	(-1/4,-1/4)

Table 2 . A Summary of Some of the Options Available to Player 1 & 2 and Their Consequences.

Example 5.1. Consider the 2-person game represented by the matrix given below.

$$\begin{array}{cc|cc|} & & & 2 \\ & & \beta_1 & \beta_2 \\ \hline & \alpha_1 & (1,2) & (-1,-4) \\ 1 & & & \\ & \alpha_2 & (-4,-1) & (2,1) \\ \hline \end{array}$$

This game has no dominated strategies and also no rationalizable set. There are 3 equilibrium points, (α_1,β_1), (α_2,β_2) and $(\frac{1}{4}\alpha_1 + \frac{3}{4}\alpha_2, \frac{3}{8}\beta_1 + \frac{5}{8}\beta_2)$. Since these are not interchangeable, the game is not Nash solvable. The minimax strategy for player 1 is $(\frac{1}{4}\alpha_1 + \frac{3}{4}\alpha_2)$ and for player 2 is $(\frac{3}{8}\beta_1 + \frac{5}{8}\beta_2)$. The maximin strategy for player 1 is $(\frac{3}{4}\alpha_1 + \frac{1}{4}\alpha_2)$ and for player 2 is $(\frac{5}{8}\beta_1 + \frac{3}{8}\beta_2)$. The maximin value for player 1 is -1/4 and for player 2 is -1/4. A summary of the various options open to player 1 and 2 and their consequences is shown in Table 2. If player 1 plays his equilibrium strategy $(\frac{1}{4}\alpha_1 + \frac{3}{4}\alpha_2)$ and player 2 plays his maximin strategy (to guarantee himself a payoff of -1/4), then player 1 gets only -1 1/4 whereas he can guarantee himself a payoff of -1/4 by playing his maximin strategy. Player 2 is in an identical situation. We let the reader judge for himself which strategy he would choose if he had to play the above game just once in the position of player 1 (or player 2) against a rational (but otherwise unknown) opponent.

ACKNOWLEDGEMENTS.

The author is grateful to John C. Harsanyi and Eric Maskin for their comments and to Stephen M. Robinson for suggesting the problem. The author alone, however, is responsible for the conclusions expressed.

REFERENCES.

1. J. C. Harsanyi, "A general theory of rational behavior in game situations," Econometrica, 34, 1966, pp. 613-634.

2. J. C. Harsanyi, "The tracing procedure: A Bayesian approach to defining a solution for n-person non-cooperative games," International Journal of Game Theory, 4, 1975, pp. 61-94.

3. J. C. Harsanyi, "A solution concept for n-person non-cooperative games," International Journal of Game Theory, 5, 1977, pp. 211-225.

4. J. C. Harsanyi, "A solution theory for non-cooperative games and its implications for cooperative games," Working Paper CP-401, Center for Research in Management Science, University of California, Berkeley, 1977.

5. J. C. Harsanyi, Rational Behavior and Bargaining Equilibrium in Games and Social Situations, Cambridge University Press, New York, 1977.

6. R. D. Luce and H. Raiffa, Games and Decisions, John Wiley & Sons, New York, 1957.

7. J. F. Nash, "Equilibrium points in n-person games," Proceedings of National Academy of Sciences, USA, 36, 1950, pp. 48-49.

8. J. F. Nash, "Non-cooperative games," Annals of Mathematics, 54, 1951, pp. 286-295 pp. 286-295.

9. J. von Neumann and O. Morgenstern, Theory of Games and Economic Behavior, Princeton University Press, New Jersey, 1953.

10. A Rapoport, Two-Person Game Theory, The University of Michigan Press, Ann Arbor, Michigan, 1966.

11. T. C. Schelling, The Strategy of Conflict, Harvard University Press, Cambridge, Massachusetts, 1960.

IMPROVEMENT IN THE COMPUTED INTERNAL ENERGIES FOR CONICAL SHAPED CHARGES

James A. Schmitt
US Army Ballistic Research Laboratory
Aberdeen Proving Ground, MD 21005

ABSTRACT. The HELP code has been used in shaped charge studies for several years. The code predicts the shape and speed of the shaped charge jet quite well. The internal energies generated by the code indicate that the jet is in a vapor-liquid state, whereas, experimental evidence suggests that the jet remains a solid. A numerical analysis of the HELP algorithm reveals the cause of the unrealistically high internal energies and a remedy. The problem is caused by the introduction of terms in the calculation of the kinetic energy which are of the order of terms neglected in this algorithm. These extraneous terms are then added to the internal energy. A noteworthy feature of the analysis and correction is that it is pertinent to many codes other than HELP. In fact, the internal energy calculation of any code based on the Particle-In-Cell algorithm, like the HELP code, can be drastically altered by these extraneous terms. The inclusion of the correction in HELP, which does not appreciably affect the running time or storage requirements of the code, results in the code predicting a solid jet with only several hot "melted" spots.

I. INTRODUCTION. Time dependent two-dimensional Eulerian computer codes like HELP[1] and HULL[2] are utilized to describe the unsteady interaction of continuous media (fluid and/or solids). Many of these continuum codes have a common ancestorial algorithm, the Particle-In-Cell method[3,4,5]. During the evolutionary process, the new codes have deviated

[1] L. J. Hageman, D. E. Wilkins, R. T. Sedgwick and J. L. Waddell, Systems Science and Software Report, TR-76-45-BK2 (1976).

[2] M. A. Fry, et al., Air Force Weapons Laboratory Report, AFWL-TR-76-183, (1976).

[3] M. W. Evans and F. H. Harlow, Los Alamos Scientific Report LA-2139,(1957).

[4] F. H. Harlow, The Particle-In-Cell Computing Method for Fluid Dynamics, in "Methods in Computational Physics", (B. Alder, S. Fernback and M. Rothenberg, eds.), Academic Press, New York, 1964.

[5] F. H. Harlow, The Particle-In-Cell Method for Numerical Solution of Problems in Fluid Dynamics, in "Proceedings of Symposia in Applied Mathematics", Vol. XV, (N. Metropolis, J. Todd, A. Tank, C. Tompkins, eds.) American Mathematical Society, Providence, RI, 1963.

substantially from the original PIC method; for example, the discrete particles were replaced by a continuum, certain Lagrangian-type features were abandoned, and for calculations in solid mechanics, material strength and effects of the deviatoric stress tensor were included. These codes can produce very successful simulations but often the results are not totally satisfactory. Such is the case with the HELP code which is used by several research laboratories and corporations for diverse applications in compressible flow and elastic-plastic flows. In certain ballistics applications at the US Army Ballistic Research Laboratory, the code predicts mass and velocities satisfactorily but an internal energy which implies a different thermodynamic state than that indicated by experimental evidence. This paper identifies the cause of the unphysical internal energies and a solution within the context of the HELP algorithm.

The internal energy algorithm in HELP is based on the finite difference approximation of the total energy equation and the kinetic energy calculated from updated mass and momentum values. This kinetic energy approximation is shown to be inaccurate in all its phases and the cause of the erroneous internal energies. The inaccuracies produce an interchange of energy at each phase which is not modelled in the governing equations. Evans and Harlow[3] identified an energy transfer mechanism in the convection phase of the original PIC method which can be seen in HELP. This mechanism is due only to spatial discretization and was illustrated in one dimension. The following analysis involves all the phases in the HELP algorithm, is two-dimensional, includes the effect of time discretization, and applies to a different code with a different energy formulation (the PIC algorithm transports total energy but directly calculates the internal energy in its other phases). Although this paper deals exclusively with the HELP algorithm, the concepts and results discussed are applicable to other codes. In particular, the same internal energy phenomena is seen in calculations performed with the HULL code.

In Section II, the governing equations which are modelled by the HELP algorithm are listed, the corresponding approximations are derived and other salient features of the algorithm are discussed. Section III contains the analysis of the kinetic energy calculation, the cause of the internal energy problem and a remedy. A numerical example illustrating the problem and the effect of the corrections is discussed in Section IV. Section V contains a summary.

II. THE HELP CODE. The unsteady motion and interaction of a continuous media can be described by a continuity equation, equations of motion, a total energy equation and an equation of state. For simplicity, we shall consider the Cartesian formulation. The appropriate two-dimensional equations in conservative form are:

$$\frac{\partial \rho}{\partial t} = -\frac{\partial}{\partial x}(\rho u) - \frac{\partial}{\partial y}(\rho v), \tag{1}$$

$$\frac{\partial(\rho u)}{\partial t} = -\frac{\partial}{\partial x}(\rho uu) - \frac{\partial}{\partial y}(\rho vu) - \frac{\partial P}{\partial x} + \frac{\partial}{\partial x}(S_{xx}) + \frac{\partial}{\partial y}(S_{xy}), \tag{2}$$

$$\frac{\partial(\rho v)}{\partial t} = -\frac{\partial}{\partial x}(\rho uv) - \frac{\partial}{\partial y}(\rho vv) - \frac{\partial P}{\partial y} + \frac{\partial}{\partial x}(S_{xy}) + \frac{\partial}{\partial y}(S_{yy}), \tag{3}$$

$$\begin{aligned}\frac{\partial(\rho E)}{\partial t} = &-\frac{\partial}{\partial x}(\rho uE) - \frac{\partial}{\partial y}(\rho vE) - \frac{\partial(uP)}{\partial x} - \frac{\partial(vP)}{\partial y} \\ &+ \frac{\partial}{\partial x}(S_{xx}u + S_{xy}v) + \frac{\partial}{\partial y}(S_{yy}v + S_{xy}u),\end{aligned} \tag{4}$$

where t, x, y, ρ, u, v, P, S_{xx}, S_{yy}, S_{xy} and E denote the time, two spatial coordinates, density, x and y components of velocity, pressure, the two normal and one shear stress components of the stress deviator tensor and specific total energy, respectively. The elements of the stress deviator tensor are functions of the velocity gradient and the pressure is computed via an equation of state of the functional form $P = P(\rho, I)$, where I is the specific internal energy. The specific internal energy is obtained as the difference of the specific total energy and the specific kinetic energy:

$$I = E - 0.5 \cdot (u^2 + v^2). \tag{5}$$

If the conservation eqs. (1) - (4) are integrated over an arbitrary control area, the time rate of change of a quantity within the control area can be related to the integrals of other quantities over the boundary enclosing that area. Performing the integration and using Green's Theorem, we obtain

$$\frac{\partial}{\partial t}\int_A \rho dA = \int_B \rho v dx - \rho u dy, \tag{6}$$

$$\frac{\partial}{\partial t}\int_A \rho u dA = \int_B \rho vudx - \rho uudy - \int_B Pdy + \int_B S_{xx}dy - S_{xy}dx, \tag{7}$$

$$\frac{\partial}{\partial t}\int_A \rho v dA = \int_B \rho vvdx - \rho uvdy + \int_B Pdx + \int_B S_{xy}dy - S_{yy}dx, \tag{8}$$

$$\begin{aligned}\frac{\partial}{\partial t}\int_A \rho E dA = \int_B \rho vEdx - \rho uEdy + \int_B Pvdx - Pudy \\ + \int_B (S_{xx}u + S_{xy}v)dy - (S_{yy}v + S_{xy}u)dx,\end{aligned} \tag{9}$$

where B is the boundary of A in the positive sense. Eq. (7), for example, equates the time rate of change of the x-component of the momentum within the area A to the product of the net mass of flow into the area and its associated specific momentum in the x-direction plus the sum of certain surface forces (the pressure and the x-components of the deviator stress tensor) exerted over the boundary enclosing the area. Such interpretations are used to determine the HELP approximations to the governing equations.

The HELP code is an Eulerian code capable of describing unsteady multi-material interaction and of treating material strength as an elastic-plastic phenomena. As a consequence of the multi-material capability, the existence and special numerical treatment of mixed cells (cells containing more than one material) are properties of the code. Although the complex treatment of the mixed cells is important and indispensable to the correct running of the code, it does make an accurate and complete analysis of the numerical treatment of mixed cells unwieldly. Consequently, the following discussion will address only the pure cell (cell containing only one material) algorithm. Furthermore, we consider only interior cells. We assume that the grid spacing Δx in the x-direction is equal as well as Δy in the y-direction. The control area A is taken to be the i^{th}, j^{th} computational cell. See Fig. 1. The left, right, top and bottom boundaries of this cell are denoted by the letters ℓ, r, a and b, respectively. A time step Δt in this explicit algorithm is determined by a Courant condition. The area integrals in eqs. (6) - (9) are approximated by $m = \rho\Delta x\Delta y$, mu, mv and mE, respectively, where m denotes the mass per unit length. All the values are at the center of the computational cell. The time derivatives are approximated by a forward difference. The values of the cell centered mass, momentum and specific total energy at the new time level are found from the

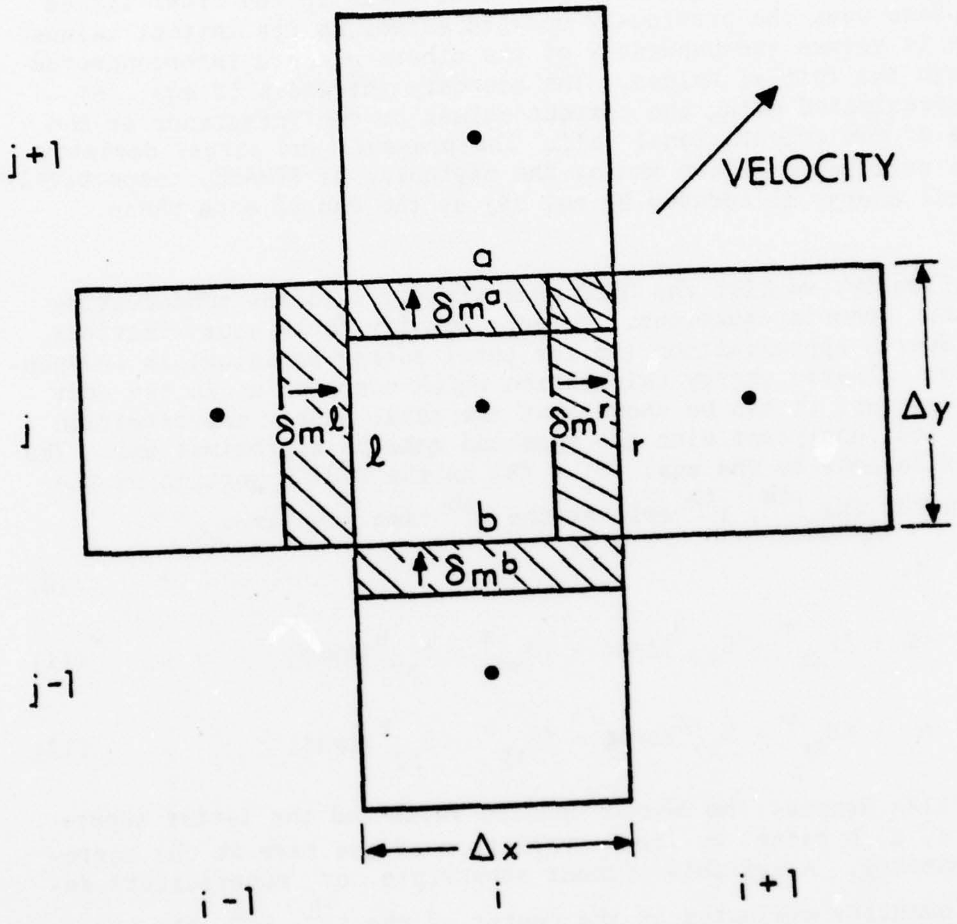


Figure 1. Computational cell for Cartesian formulation of HELP.

values at the previous time level. This is accomplished in three stages by determining the time rate of change of the mass, momentum and total energy due to i) the effects of the deviatoric stresses, ii) the effects of the pressure, and iii) the effects of the convection terms. These phases are appropriately named SPHASE, HPHASE and TPHASE, respectively. During each time step, each value of the mass, momentum and total energy is updated sequentially by each phase in the order listed and each phase uses the previously updated values as its initial values. Each phase is solved independently of the others and are interconnected only through the initial values. The boundary integrals in eqs. (6) - (9) are approximated using the current values of the integrands at the boundaries of the computational cell. The pressure and stress deviator tensor are calculated before and at the beginning of SPHASE, respectively. The internal energy is updated by eq. (5) at the end of each phase.

Specifically, we list the HELP approximations to the conservation of mass and momenta equations. We only specify these approximations, since the fourth approximation (to the total energy equation) is independent of the kinetic energy calculation which concerns us in the next section. However, it can be shown that the total energy approximation is correct and consistent with the mass and momenta approximations. The HELP approximations to the eqs. (6) - (8) in the SPHASE portion of the calculation for the i^{th}, j^{th} cell at the n^{th} time step are:

$$\tilde{m} = m, \tag{10}$$

$$\widetilde{mu} = mu + (S_{xx}{}^{r} - S_{xx}{}^{\ell})\Delta y\Delta t + (S_{xy}{}^{a} - S_{xy}{}^{b})\Delta x\Delta t, \tag{11}$$

$$\widetilde{mv} = mv + (S_{xy}{}^{r} - S_{xy}{}^{\ell})\Delta y\Delta t + (S_{yy}{}^{a} - S_{yy}{}^{b})\Delta x\Delta t, \tag{12}$$

where the tilde denotes the SPHASE updated value and the letter superscripts r, ℓ, a, b refer to the evaluation of the term at the corresponding boundary. A variable without subscripts or superscripts denotes that quantity evaluated at the center of the i^{th}, j^{th} cell at the n^{th} time level. The boundary value of a variable is the average of its cell centered values adjacent to that boundary, for example, $S^{r}_{xx} = 0.5$ $[(S_{xx})^{n}_{i+1,j} + (S_{xx})^{n}_{i,j}]$, where $(S_{xx})^{n}_{i,j} = S_{xx}[(i-\frac{1}{2})\Delta x,(j-\frac{1}{2})\Delta y,n\Delta t]$. See Fig. 1. The approximations (10) - (12) can be derived from a physical interpretation of the SPHASE portions of eqs. (6) - (8). For example, consider approximation (11). The effect of the stress deviator tensor on the time rate of change of the x-component of the momentum $[(\widetilde{mu}) - (mu)]/\Delta t$ during the entire time step Δt is governed by the stress elements S_{xx} and S_{xy} on the right and left and top and bottom boundaries, respectively. Furthermore, within the context of SPHASE (assuming the post SPHASE values are those at the end of the time step), approximations (11) and (12) can be shown to be first order in time and second order in space to the relevant portion of eqs. (2) - (3), respectively. To determine the order of the approximation (11), for example, we substitute the values of the mass and boundary stresses into approximation (11) and obtain

$$\frac{(\widetilde{\rho u}) - (\rho u)}{\Delta t} = \frac{(S_{xx})^n_{i+1,j} - (S_{xx})^n_{i-1,j}}{2\Delta x} + \frac{(S_{xy})^n_{i,j+1} - (S_{xy})^n_{i,j-1}}{2\Delta y} . \tag{13}$$

By a simple Taylor series expansion about the center of the i^{th}, j^{th} cell at the n^{th} time level, eq. (13) can be shown to exhibit the stated properties.

The HPHASE approximations are

$$\overline{m} = m, \tag{14}$$

$$\overline{mu} = m\tilde{u} - (P^r - P^\ell)\Delta y\Delta t, \tag{15}$$

$$\overline{mv} = m\tilde{v} - (P^a - P^b)\Delta x\Delta t, \tag{16}$$

where the bar denotes the HPHASE updated value and $P = P(\rho, I)$. Similar comments to those on the SPHASE approximations hold true for HPHASE.

For simplicity in the discussion of the TPHASE approximations, we assume that the velocity has both positive x and y components. The TPHASE approximations which model the convection between cells are:

$$m^{n+1} = m + \delta m^\ell - \delta m^r + \delta m^b - \delta m^a, \tag{17}$$

$$(mu)^{n+1} = m\bar{u} + \delta m^\ell \bar{u}_{i-1,j} - \delta m^r \bar{u} + \delta m^b \bar{u}_{i,j-1} - \delta m^a \bar{u}, \tag{18}$$

$$(mv)^{n+1} = m\bar{v} + \delta m^\ell \bar{v}_{i-1,j} - \delta m^r \bar{v} + \delta m^b \bar{v}_{i,j-1} - \delta m^a \bar{v}, \tag{19}$$

where δm^ℓ, δm^b, δm^r, δm^a denote the convected mass per unit length from the left and bottom cell and to the right and top, respectively. See Fig. 1. In general, $\delta m^d = \rho^d L^d u^d \Delta t$, where ρ^d denotes the density of the cell from which the mass is transported, u^d is the interpolated value of the velocity component normal to the cell boundary and L^d is the length of the cell boundary through which the mass is moved. For example, the factors in δm^ℓ are $\rho^\ell = \rho_{i-1,j}$, $u^\ell = 0.5\ (\bar{u}+\bar{u}_{i-1,j})/[1+\Delta t(\bar{u}_{i-1,j} - \bar{u})/\Delta x]$ and $L^d = \Delta y$. We note that u^d represents the transport velocity of δm^ℓ based on linear approximations over the time step Δt. The intuitive explanation of the TPHASE approximations for eq. (17) is that the mass at the end of TPHASE (the final value at (n+1) time level) is the mass originally in the cell plus the mass transported into the cell (δm^ℓ, δm^b) minus the mass transported from the cell (δm^r, δm^a). For the momentum equations, a similar situation exists except that now each convected mass is associated with the specific momentum of its "donor" cell. Within the context of TPHASE (assuming the post HPHASE values are the initial values at the n^{th} time level), the approximations (17) - (19) can be shown to be first order in time and space to the TPHASE portions of eqs. (1) - (3), respectively. By substituting the appropriate approximation for the masses, eq. (17), for example, can be rewritten as

$$\frac{\rho^{n+1} - \rho}{\Delta t} = - \frac{\rho u^r - \rho_{i-1,j} u^\ell}{\Delta x} - \frac{\rho v^a - \rho_{i,j-1} v^b}{\Delta y} \tag{20}$$

A Taylor Series expansion about the center of the i^{th}, j^{th} cell at the "n^{th} time level" of each term, enables eq. (20) to be rewritten as

$$\frac{\partial \rho}{\partial t} + O(\Delta t) = - \frac{\partial (\rho u)}{\partial x} - \frac{\partial (\rho v)}{\partial y} + O(\Delta x) + O(\Delta y) + O(\Delta t) \tag{21}$$

which establishes the assertion.

III. <u>KINETIC ENERGY CALCULATION</u>. The partial differential equation governing kinetic energy can be derived from eqs. (1) - (3) by the following identity:

$$\frac{\partial}{\partial t} (\rho e) = u \frac{\partial \rho u}{\partial t} + v \frac{\partial \rho v}{\partial t} - \frac{1}{2} (u^2+v^2) \frac{\partial \rho}{\partial t}, \tag{22}$$

and can be written as

$$\begin{aligned} \frac{\partial}{\partial t} (\rho e) = & - \frac{\partial}{\partial x} (\rho u e) - \frac{\partial}{\partial y} (\rho v e) - u \frac{\partial P}{\partial x} - v \frac{\partial P}{\partial y} \\ & + u\left(\frac{\partial S_{xx}}{\partial x} + \frac{\partial S_{xy}}{\partial y}\right) + v\left(\frac{\partial S_{xy}}{\partial x} + \frac{\partial S_{yy}}{\partial y}\right), \end{aligned} \tag{23}$$

where $e = 0.5\ (u^2 + v^2)$. An interpretation of the above manipulation is that given the exact solutions of eqs. (1) - (3), the derived function e is identical to the exact solution of eq. (23). We shall now show that the HELP approximations do not share this property.

The specific kinetic energy e at the end of each phase is computed via

$$e = 0.5 \left\{ [(mu)/m]^2 + [(mv)/m]^2 \right\} \tag{24}$$

using the updated values of the mass and momenta from that phase. By using the approximations (10) - (12), (14) - (16) and (17) - (19), eq. (24) and the expressions for the mass in terms of the density, we can write the implicit formulas used to determine the updated specific kinetic energy at the different phases in terms of the initial values at SPHASE, HPHASE and TPHASE. The resulting expressions are termed implicit, since they are never explicitly used to calculate the kinetic energy but are numerically equivalent to eq. (24). The results are:

$$\begin{aligned}\frac{(\widetilde{\rho e})-\rho e}{\Delta t} &= u\left[\frac{S_{xx}{}^{r}-S_{xx}{}^{\ell}}{\Delta x} + \frac{S_{xy}{}^{a}-S_{xy}{}^{b}}{\Delta y}\right] + v\left[\frac{S_{xy}{}^{r}-S_{xy}{}^{\ell}}{\Delta x} + \frac{S_{yy}{}^{a}-S_{yy}{}^{b}}{\Delta y}\right] \\ &+ \left\{\frac{\Delta t}{2\rho}\left[\left(\frac{S_{xx}{}^{r}-S_{xx}{}^{\ell}}{\Delta x} + \frac{S_{xy}{}^{a}-S_{xy}{}^{b}}{\Delta y}\right)^2 + \left(\frac{S_{xy}{}^{r}-S_{xy}{}^{\ell}}{\Delta x} + \frac{S_{yy}{}^{a}-S_{yy}{}^{b}}{\Delta y}\right)^2\right]\right\},\end{aligned} \tag{25}$$

$$\begin{aligned}\frac{(\bar{\rho e}) - (\widetilde{\rho e})}{\Delta t} &= -\tilde{u}\left[\frac{p^r - p^\ell}{\Delta x}\right] - \tilde{v}\left[\frac{p^a - p^b}{\Delta y}\right] \\ &+ \left\{\frac{\Delta t}{2\rho}\left[\left(\frac{p^r - p^\ell}{\Delta x}\right)^2 + \left(\frac{p^a - p^b}{\Delta y}\right)^2\right]\right\},\end{aligned} \tag{26}$$

$$
\begin{aligned}
\frac{(\rho e)^{n+1} - (\overline{\rho e})}{\Delta t} &= - \frac{\rho u^{r}\bar{e} - \rho_{i-1,j}u^{\ell}\bar{e}_{i-1,j}}{\Delta x} - \frac{\rho v^{a}\bar{e} - \rho_{i,j-1}v^{b}\bar{e}_{i,j-1}}{\Delta y} \\
&- \Bigg\{ \frac{\rho_{i-1,j}\rho}{2\rho^{n+1}} u^{\ell}\left[\Delta x - u^{r}\Delta t - v^{a}\Delta t \frac{\Delta x}{\Delta y}\right]\left[\left(\frac{\bar{u}-\bar{u}_{i-1,j}}{\Delta x}\right)^{2} + \left(\frac{\bar{v}-\bar{v}_{i-1,j}}{\Delta x}\right)^{2}\right] \\
&+ \frac{\rho_{i,j-1}\rho v^{b}}{2\rho^{n+1}} \left[\Delta y - v^{a}\Delta t - u^{r}\Delta t \frac{\Delta x}{\Delta y}\right]\left[\left(\frac{\bar{u}-\bar{u}_{i,j-1}}{\Delta y}\right)^{2} + \left(\frac{\bar{v}-\bar{v}_{i,j-1}}{\Delta y}\right)^{2}\right] \qquad (27) \\
&+ \frac{\rho_{i-1,j}\rho_{i,j-1}v^{b}u^{\ell}}{2\rho^{n+1}} \Delta t \frac{\Delta x}{\Delta y}\left[\left(\frac{\Delta y}{\Delta x}\frac{\bar{u}-\bar{u}_{i,j-1}}{\Delta y} - \frac{\bar{u}-\bar{u}_{i-1,j}}{\Delta x}\right)^{2}\right. \\
&+ \left.\left(\frac{\Delta y}{\Delta x}\frac{\bar{v}-\bar{v}_{i,j-1}}{\Delta y} - \frac{\bar{v}-\bar{v}_{i-1,j}}{\Delta x}\right)^{2}\right] \Bigg\} .
\end{aligned}
$$

By a Taylor series analysis similar to those performed in Section II, it can be shown that within the context of SPHASE and HPHASE, approximations (25) and (26) are first order in time and second order in space to the SPHASE and HPHASE portion of eq. (23), respectively, and that within the context of TPHASE, approximation (27) is first order in both time and space to the TPHASE portion of eq. (23). See Ref. 6. The order of the kinetic energy approximation is in accord with the other approximations in the three phases. However, each approximation (25) - (27) includes terms that are of the order of the formal truncation error: terms of order Δt is SPHASE and HPHASE and order Δt, Δx and Δy in TPHASE. These terms are enclosed by braces. These extraneous terms with respect to the truncation error are consequences of calculating the kinetic energy from the updated values of the mass and momenta equations in each phase. Furthermore, these terms do not model any term of the kinetic energy equation. In fact, if one would write directly a finite difference approximation to eq. (23) in a consistent manner with the HELP approximations of eqs. (1) - (3), the result would be eqs. (25) - (27) without the first order terms. Although in the theoretical limit as the mesh approaches zero the two approximations are the same, in practice the inclusion of terms of the order of the truncation error alters the computed value of the kinetic energy of a cell at each cycle. This alteration is accumulative. In a calculation involving hundreds of cycles, the final kinetic energy may differ by more than an order of magnitude.

[6] J. Lacetera, Jr., J. M. Lacetera and J. A. Schmitt, US Army Ballistic Research Laboratory Report (in preparation).

The coefficients of the Δt terms in all the three phases are positive and consequently, these terms increase the calculated kinetic energy. The coefficient of the spatial increments Δx and Δy appearing only in TPHASE are negative and decrease the calculated kinetic energy. Furthermore, it can be shown that for equal spatial meshes ($\Delta x = \Delta y$) the entire first order term in the TPHASE calculation is negative for Courant number less than a half and decreases the kinetic energy. A common feature of all these terms is their quadratic dependence on the first spatial derivatives: in SPHASE of the deviator stress elements, in HPHASE of the pressure and in TPHASE of the velocity components. In regions of small gradients, the contribution of these terms of the kinetic energy value will be small. However, in regions of large gradients, these terms can contribute substantially to the calculated kinetic energy even though they are formally of the order of the truncation error. In an application discussed in the next section, these terms can account for as much as 15% of the calculated kinetic energy within a cell during one time step.

The effects of these terms are not confined to the kinetic energy calculation. It can be shown that the specific total energy is calculated in a manner consistent with that of the mass and momenta equations. Thus, the effect of the extraneous terms in the kinetic energy calculation is directly translated to the internal energy calculation via eq. (5) with a reverse effect. The order Δt terms decrease the internal energy and the order Δx and Δy terms increase it. Thus, these terms can be interpreted as a transfer mechanism which is not modeled by the governing equations and which converts internal energy into kinetic energy and kinetic energy into internal energy. Consider, for example, the one dimensional first order energy approximation in TPHASE for motion in the x-direction. The only first order term that the internal energy calculation includes is the positive term:

$$\frac{\rho_{i-1}\rho u^{\ell}}{2\rho^{n+1}} (\Delta x - u^{r}\Delta t) \left(\frac{\bar{u} - \bar{u}_{i-1,j}}{\Delta x}\right)^2 \tag{28}$$

Expanding expression (28) in a Taylor series about the cell center and the n^{th} time level, we obtain

$$(\lambda + \lambda')\left[\frac{\partial u}{\partial x}\right]^2 \tag{29}$$

to the lowest order, where $\lambda = 0.5\ u\Delta x$ and $\lambda' = -0.5\ u^2\Delta t$. If $\lambda + \lambda'$ were the coefficient of viscosity, then expression (29) would be identical to the viscosity term in the one dimensional internal energy equation. Thus, the energy transfer mechanism in this case is an explicit artificial viscosity term. Evans and Harlow[3] identified the term corresponding to $\lambda[\partial u/\partial x]^2$ in their one dimensional analysis of the original PIC code.

The term $\lambda'[\partial u/\partial x]^2$ is not included in their analysis, since they did not include the effect of time differencing. From expression (29), we see that the time discretization decreases the amount of kinetic energy converted to internal energy in TPHASE. This explicit type of artificial viscosity is confined only to the TPHASE energy calculation and is, in addition to the implicit artificial viscosity, already inherent in a first order algorithm[7].

Irrespective of the algebraic sign of the first order terms in eqs. (25) - (27), they do alter not only the internal energy value but also the pressure via the equation of state, the strength calculation via the internal energy dependent yield criteria, and consequently, the entire calculation. To determine the effects of omitting these terms in a calculation, the HELP code was modified to allow a kinetic energy calculation which did not include the first order terms present in eqs. (25) - (27). Consequently, the kinetic energy was not computed by the updated mass and momentum values but was considered a separate dependent variable. This kinetic energy was updated according to a direct finite differencing of eq. (23) in a manner consistent with the other approximations and was stored in an array. The results of this modified formulation are compared to those of the original version for a typical HELP application at the Ballistic Research Laboratory in the next section.

IV. EXAMPLE. The HELP code is used primarily at the Ballistic Reseearch Laboratory to simulate the detonation and jet formation of an armor piercing warhead called a shaped charge. We will consider an unconfined conical shaped charge (see Fig. 2a). The actual warhead is obtained by rotating Fig. 2a about the axis of symmetry. The explosive is detonated and the detonation wave collapses the conical liner towards the axis of symmetry with a varying velocity. Sixteen microseconds after detonation, the liner consists of three parts (Fig. 2b): The undeformed liner, the low velocity large mass slug and the high velocity small mass jet. By performing a Galilean transformation at the point where an original ring of liner impinges on the axis of symmetry, a stagnation point in the flow can be shown to exist. This stagnation point divides the deformed liner into the slug and jet. It is the jet which pierces the armor and is of prime concern to the shaped charge designer. By assuming that the liner can be modeled as an incompressible material under typical loading conditions, several researchers[8,9,10] have developed

[7] P. J. Roache, "Computational Fluid Dynamics", Chapter V, Hermosa, Albuquerque, 1976.

[8] G. Birkhoff, D. P. MacDougall, E. M. Pugh and G. Taylor, J. Appl. Phys. 19, (1948), 563.

[9] E. M. Pugh, R. J. Eichelberger and N. Rostoker, J. Appl. Phys. 23 (1952), 532.

[10] A. R. Kiwan and H. Wisniewski, BRL Report No. 1620 (1972).

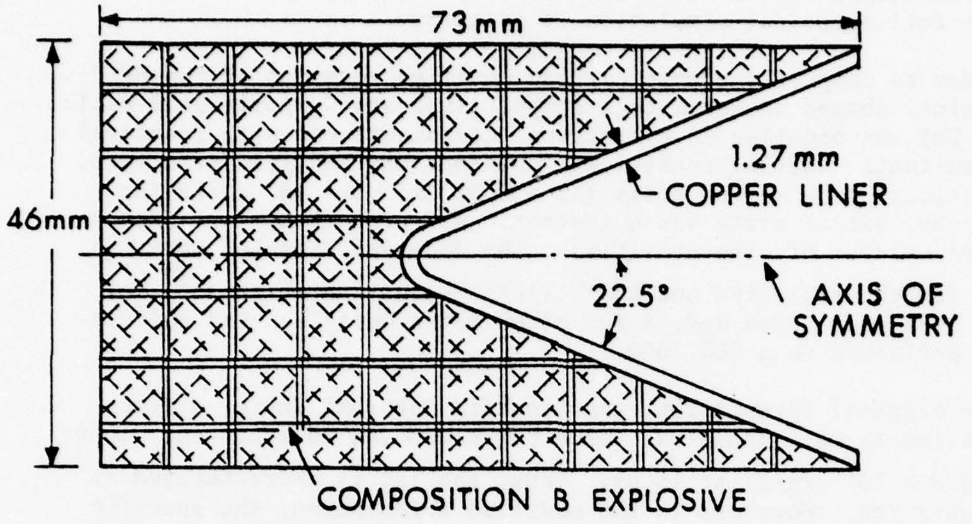


Figure 2a. Initial conical shaped charge configuration.

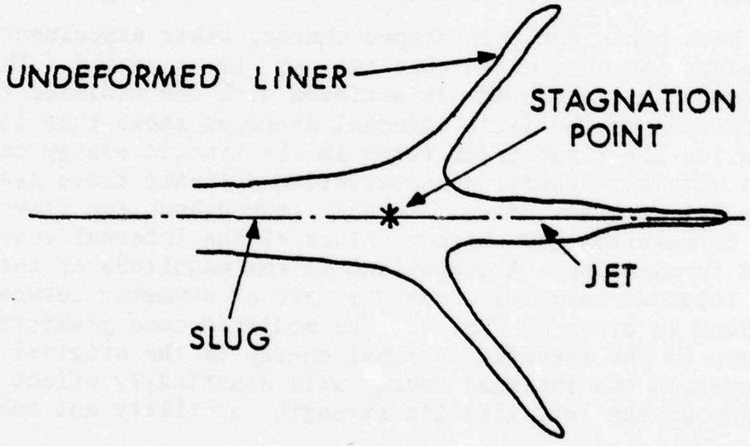


Figure 2b. Copper configuration at 16 μs for conical shaped charge.

analytical or simple numerical models to determine the velocity field. However, when compressibility, strength, and thermodynamic effects are included, a full numerical simulation is necessary.

In order to compare the original and modified versions of the HELP code, a conical shaped charge with a copper liner and Composition B explosive (Fig. 2a) was modelled in cylindrical coordinate and with identical material constants, initial conditions, code options and grid structure. The computational mesh was 60 cells (Δr = 0.52 mm) by 187 cells (Δz = 0.52 mm). The initial state was quiescent: all properties zero except the standard values of the densities. The Jones-Wilkins-Lee equation of state[11] for the explosive and the Tillotson equation of state[12] for the copper were used. See Ref. 6 for other setup details. The calculations were performed on a CDC 7600.

In the original formulation, the specific internal energy of each cell in the jet is well above the value corresponding to incipient vaporization (13.8 x 10^9 erg/g) at 16 μs. Thus, the jet is characterized as a vapor-liquid jet. However, in the modified formulation, the specific internal energy of the same cells is generally below that for melt (5.3 x 10^9 erg/g) and a solid jet with several melted sections is predicted. See Fig. 3. The specific internal energy values which correspond to the melting point and incipient vaporization point of copper are contained within the Tillotson equation of state. Although no actual temperature has been taken for this shaped charge, other experimental evidence[13,14] supports the conclusion that the jet is a solid. Thus, qualitative thermodynamic agreement is achieved with the modified formulation. This comparison of the jet's internal energies shows that in the original formulation the first order terms in the kinetic energy calculation associated with the spatial discretization dominate those associated with the temporal discretization. In fact, throughout the flow field, the original formulation gave higher values of the internal energy than the modified formulation. A comparison of the magnitude of the computed specific internal energies along the axis of symmetry between the two formulations is given in Fig. 4. The modified code predicts up to an 88% decrease in the specific internal energy of the original code. Such major changes in the internal energy will drastically effect the material properties of the jet, like its strength, ductility and cohesion. Two other

[11] E. Lee, M. Finger and W. Collins, Lawrence Livermore Laboratory Report UCID-16189 (1973).

[12] J. H. Tillotson, General Atomic Report GA-3216 (1962).

[13] W. G. Von Holle and J. J. Trimble, US Army Ballistic Research Laboratory Report 2624 (1976).

[14] W. G. Von Holle and J. J. Trimble, US Army Ballistic Research Laboratory Report 2004 (1977).

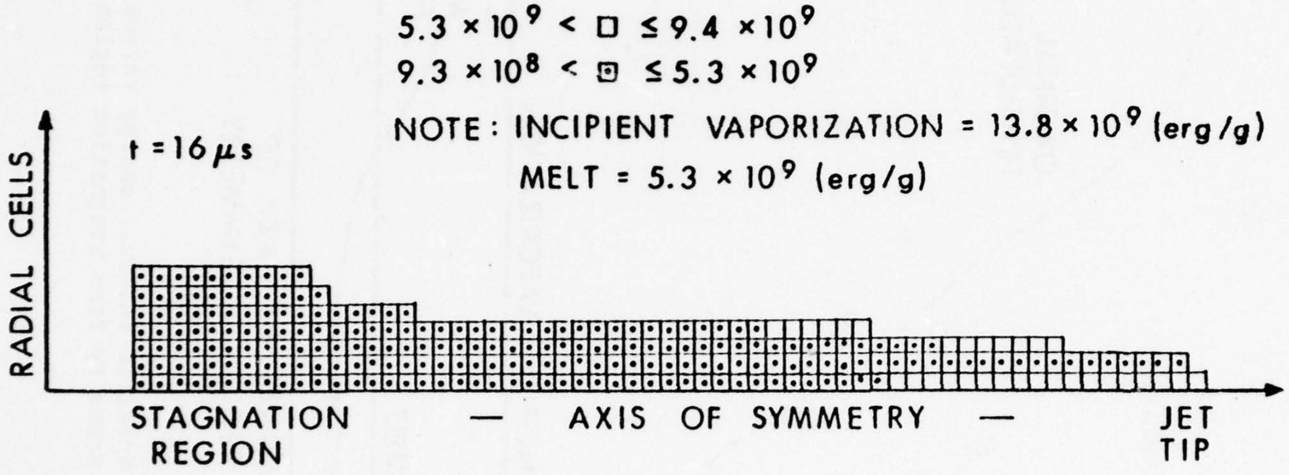


Figure 3. Cell values of specific internal energy (erg/g) of copper jet computed by modified formulation.

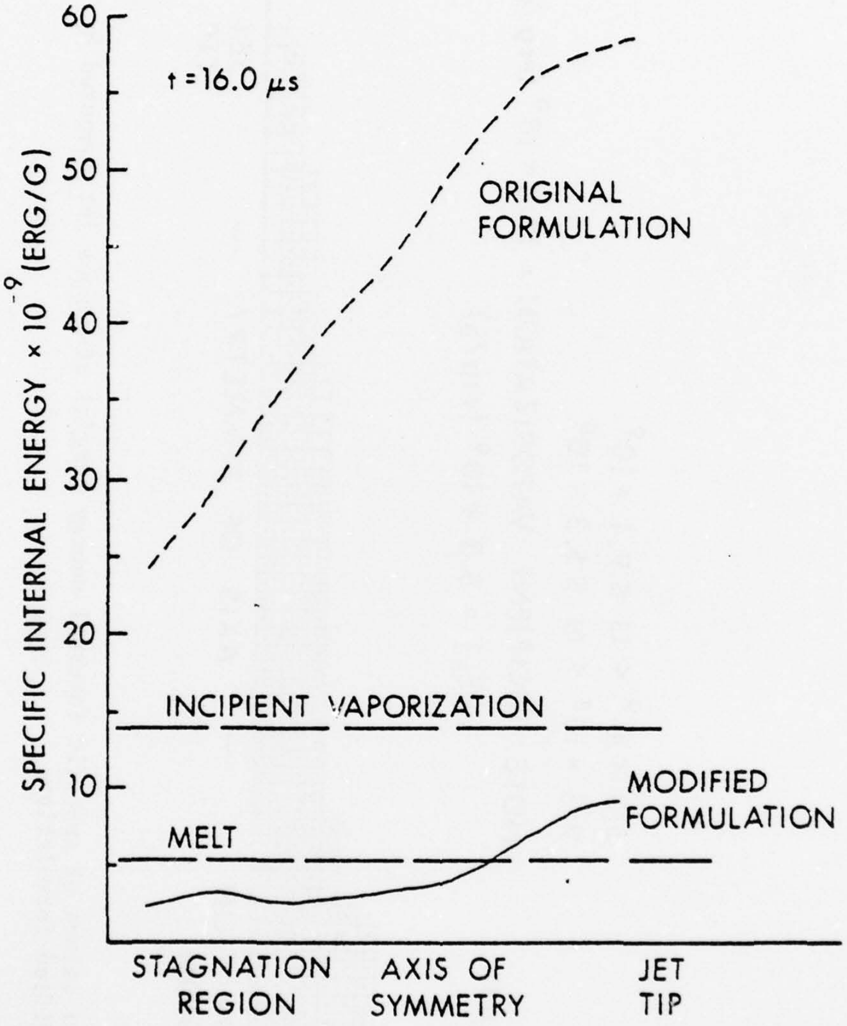


Figure 4. Specific internal energy versus distance along axis of symmetry from stagnation region to the jet tip at t = 16 μs.

important quantities to the warhead designers are the density and axial velocity. Figs. 5 and 6 show comparisons of the compression and axial velocity of the jet along the axis of symmetry. The jet density of the modified code is approximately 10% higher than the original. Near the stagnation region, the modified formulation gives an improved result: a compression ($\rho > \rho_o$) of the copper. The density discontinuity near the jet tip which may not be physical is of smaller magnitude in the modified version. The axial velocities (Fig. 6) are virtually identical except near the jet tip where the physical inverse velocity gradients are present but have small magnitudes. The larger axial velocity values of the original version caused the slightly longer jet (approximately two computational cells). The relative jet tip velocity (jet tip velocity minus slug end velocity) of 6.02×10^5cm/s is 7.9% lower than that indicated by experimental flash radiographs. The discrepancies between the density and axial velocities would be much greater if the code did not have an artificial cut off value (13.8×10^9 erg/g) for the specific internal energy in the equation of state for the liner material. Since the cell values of the jet's specific internal energy are well above the cut-off value in the original formulation, the pressure, velocity, total energy and mass do not include the effects of the computed high internal energies.

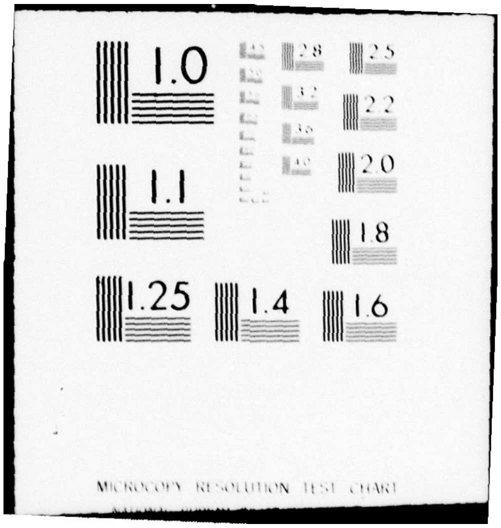
V. SUMMARY. We have shown that inclusion of terms which are of the order of the truncation error in an approximation can severely alter a calculation. Although in certain computations these extraneous terms may remain negligible, in others they can be significant and produce spurious results. A case in point has been cited with the HELP code and its kinetic energy calculation, and the effects illustrated by an application to a problem in warhead mechanics. In the original HELP code, the updated values of kinetic energy were computed as consequences of the updated mass and momenta values. This value is shown to deviate from that computed directly by a first order approximation of the kinetic energy by first order terms which depend quadratically on the spatial derivatives of the velocity, pressure and elements of the deviator stress tensor. These terms become significant when applied to problems involving large gradients; such as conical shaped charge simulations. A method was suggested to avoid these terms within the confines of the basic HELP algorithm. The modified code predicted significantly better internal energies. In other applications, the prognosis for the modified formulation is good, since the approximation to the kinetic energy is more accurately calculated.

In the TPHASE section of the original HELP algorithm the extraneous terms were identified with an explicit artificial viscosity in the internal energy calculation. Consequently, the modified formulation may require implementation of the artificial viscosity option available in cases where the original formulation did not.

A noteworthy feature of the analysis and modification is that it is directly relevant to codes other than HELP. In fact, any code with a HELP type algorithm can be erroneously affected by extraneous terms in the kinetic energy calculation. In particular, the same unphysical internal energies in the jet are computed by the HULL code.

AD-A064 433 ARMY RESEARCH OFFICE RESEARCH TRIANGLE PARK N C F/G 12/1
TRANSACTIONS OF THE CONFERENCE OF ARMY MATHEMATICIANS(24TH) HEL--ETC(U)
JAN 79
UNCLASSIFIED ARO-79-1 NL

2 OF 6
AD
A064433


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
MICROCOPY RESOLUTION TEST CHART

important quantities to the warhead designers are the density and axial velocity. Figs. 5 and 6 show comparisons of the compression and axial velocity of the jet along the axis of symmetry. The jet density of the modified code is approximately 10% higher than the original. Near the stagnation region, the modified formulation gives an improved result: a compression ($\rho > \rho_o$) of the copper. The density discontinuity near the jet tip which may not be physical is of smaller magnitude in the modified version. The axial velocities (Fig. 6) are virtually identical except near the jet tip where the physical inverse velocity gradients are present but have small magnitudes. The larger axial velocity values of the original version caused the slightly longer jet (approximately two computational cells). The relative jet tip velocity (jet tip velocity minus slug end velocity) of 6.02×10^5 cm/s is 7.9% lower than that indicated by experimental flash radiographs. The discrepancies between the density and axial velocities would be much greater if the code did not have an artificial cut off value (13.8×10^9 erg/g) for the specific internal energy in the equation of state for the liner material. Since the cell values of the jet's specific internal energy are well above the cut-off value in the original formulation, the pressure, velocity, total energy and mass do not include the effects of the computed high internal energies.

V. SUMMARY. We have shown that inclusion of terms which are of the order of the truncation error in an approximation can severely alter a calculation. Although in certain computations these extraneous terms may remain negligible, in others they can be significant and produce spurious results. A case in point has been cited with the HELP code and its kinetic energy calculation, and the effects illustrated by an application to a problem in warhead mechanics. In the original HELP code, the updated values of kinetic energy were computed as consequences of the updated mass and momenta values. This value is shown to deviate from that computed directly by a first order approximation of the kinetic energy by first order terms which depend quadratically on the spatial derivatives of the velocity, pressure and elements of the deviator stress tensor. These terms become significant when applied to problems involving large gradients; such as conical shaped charge simulations. A method was suggested to avoid these terms within the confines of the basic HELP algorithm. The modified code predicted significantly better internal energies. In other applications, the prognosis for the modified formulation is good, since the approximation to the kinetic energy is more accurately calculated.

In the TPHASE section of the original HELP algorithm the extraneous terms were identified with an explicit artificial viscosity in the internal energy calculation. Consequently, the modified formulation may require implementation of the artificial viscosity option available in cases where the original formulation did not.

A noteworthy feature of the analysis and modification is that it is directly relevant to codes other than HELP. In fact, any code with a HELP type algorithm can be erroneously affected by extraneous terms in the kinetic energy calculation. In particular, the same unphysical internal energies in the jet are computed by the HULL code.

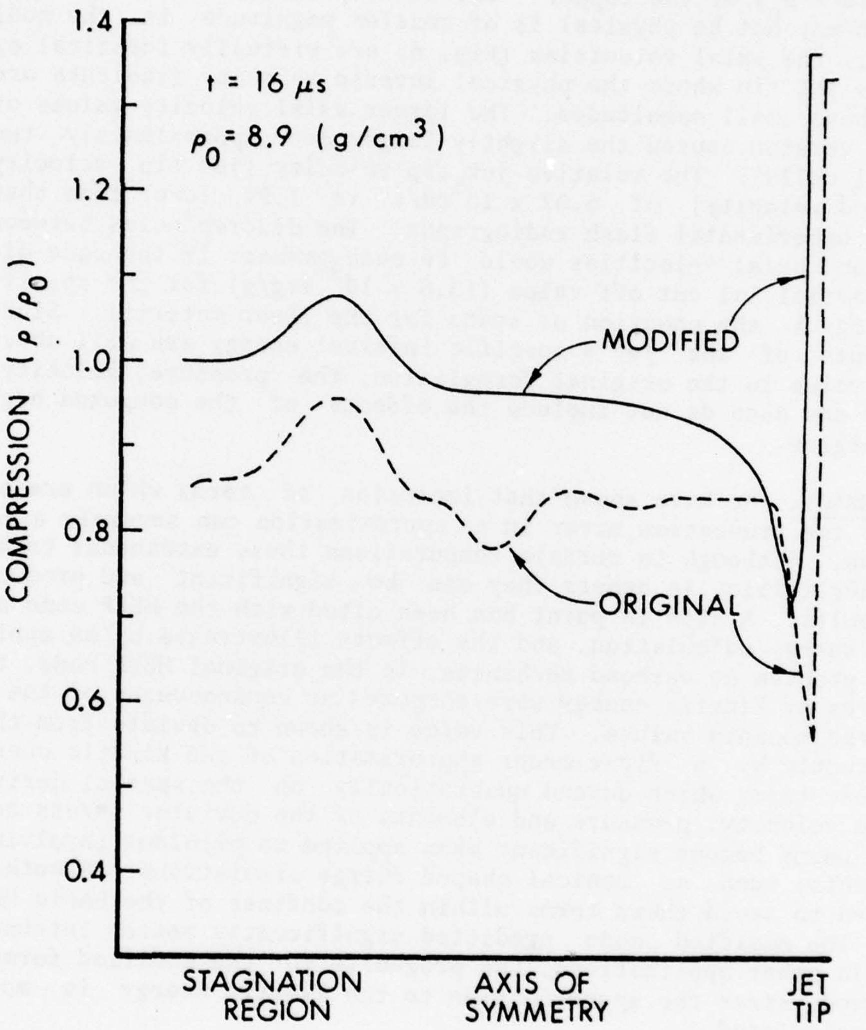


Figure 5. Compression versus distance along axis of symmetry from stagnation region to the jet tip at 16 μs.

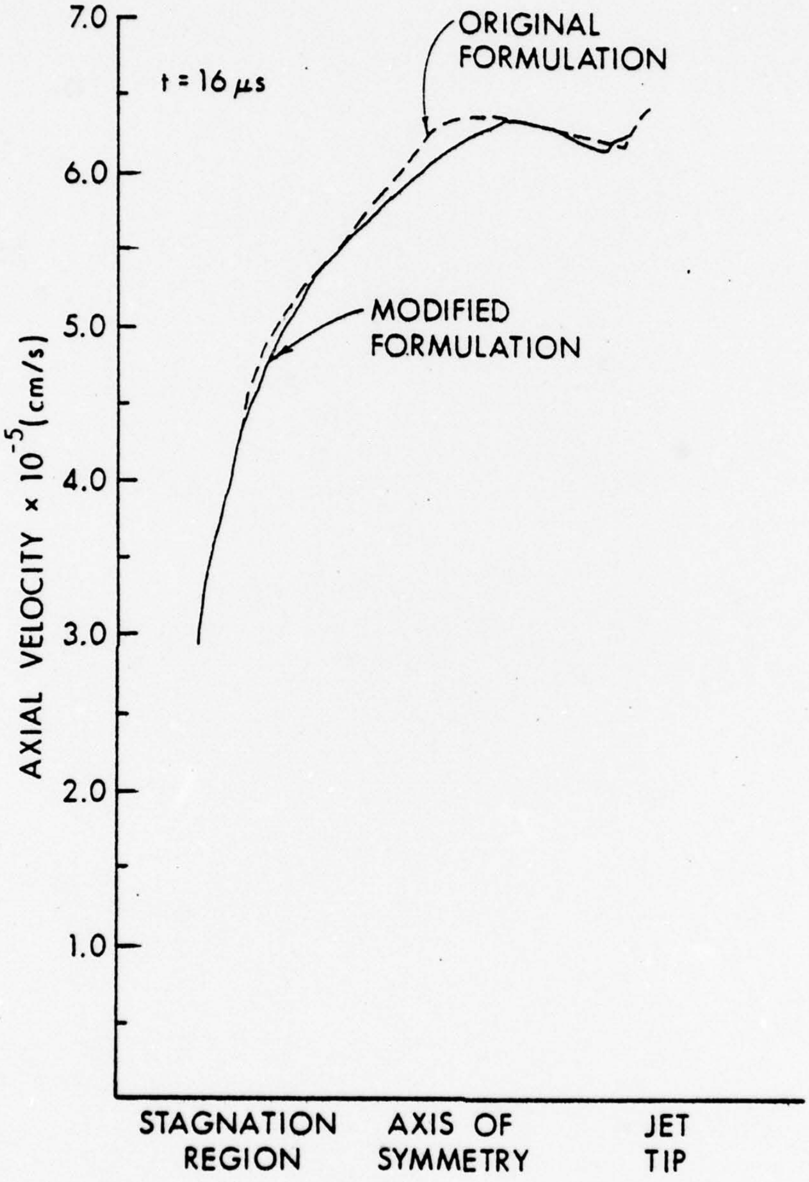


Figure 6. Axial velocity versus distance along axis of symmetry from stagnation region to the jet tip at 16 μs.

STRESS INTENSITY FACTORS FOR A CIRCULAR RING WITH UNIFORM ARRAY OF RADIAL CRACKS USING CUBIC ISOPARAMETRIC SINGULAR ELEMENTS

S. L. Pu and M. A. Hussain
U.S. Army Armament Research and Development Command
Benet Weapons Laboratory
Watervliet Arsenal, Watervliet, NY 12189

ABSTRACT. The plane problem of a uniform array of equal depth radial cracks originating at the internal boundary of a pressurized circular ring is considered. The finite element method using 12-node quadrilateral, isoparametric elements is adopted. The collapsed singular elements recently developed by the authors are used around the crack tip. The stress intensity factors at a crack tip can be obtained for any finite number of radial cracks and for a large variety of diameter ratios and crack-depth to wall-thickness ratios.

For the special case of a single radial crack and two diametrically opposed radial cracks, stress intensity factors have been obtained by Bowie and Freese using modified mapping-collocation method and by Shannon using a very large number of constant-strain triangular elements. Results of these two different approaches agree quite well except for shallow cracks relative to the cylinder wall thickness. The present finite element results using a maximum of seventeen elements are in better agreement with those of Bowie and Freese, including the results for shallow cracks.

For the case of forty radial cracks in a cylinder of diameter ratio 2.0, Goldthorpe obtained an empirical formula for the stress intensity factor based on an approximate procedure applied to data of Tweed and Rooke for forty radial cracks from a hole in an infinite plate. The present results agree with Goldthorpe's results for shallow cracks. For large crack-depth to wall-thickness ratios, Goldthorpe's formula tends to be too low for the stress intensity factors.

The current study has shown that the ring with two diametrically opposed cracks is in general the weakest configuration (highest value in K_I). In the range of diameter ratio 1.5 to 2.5, the ring with three radial cracks is also weaker than that with only one radial crack. For more than three cracks, the denser the radial cracks are the more stable the ring will be.

I. INTRODUCTION.

The plane problem of radial cracks, equal and finite in length, originating at a circular hole in an infinite plate under uniaxial or biaxial tension has been solved by Bowie [1]. His solution is based on the complex variable method. He obtained numerical results for a single crack and two diametrically opposite cracks. Using this analysis, Kutter [2] computed the stress distribution for a maximum of sixteen radial cracks. A special case when the radius of the circular hole is zero was considered by Westmann [3]. His solution is obtained utilizing Mellin transforms. The crack tip stress intensity factors are numerically evaluated for any number of radial cracks. Using Mellin transforms, Tweed and Rooke [4-6] reconsidered Bowie's and Westmann's problems. They obtained stress intensity factors for various crack numbers and lengths. A problem of this nature is of particular interest in the field of rock mechanics such as underground cavities under hydraulic pressure or blasting. It is also useful for circular cutouts in sheet material. The application of this problem to axial cracks in a circular hollow cylinder is of limited practical value because the solution is only valid for very large diameter ratios.

Various approximate techniques have been developed for the stress intensity factor of radial cracks radiating at the bore of a circular ring. Winnie and Wundt [7] used Bowie's infinite plate solution and a mean-stress concept to obtain strain-energy release rate for a bored and notched rotating disk of large diameter ratios. Williams and Isherwood [8] presented an approximate method for the calculation of the strain-energy release rate for finite plates. They obtained numerical results for rotating discs with diameter ratios less than that required in [7]. In the study of cracked gun barrels, Goldthorpe [9] computed the stress intensity factor for a pressurized cylinder of diameter ratio 2.0 with forty radial cracks based on an approximate procedure by Cartright and Rooke [10] and the results of radial cracks in an infinite plate by Tweed and Rooke [6]. The modified mapping-collocation technique which combines modified versions of conformal mapping and boundary collocation methods was originally presented by Bowie and Neal [11] as a procedure for the analysis of an internal crack in a finite geometry. This technique was employed by Bowie and Freese [12] to evaluate stress intensity factors for a circular ring with a single radial crack at the inner hole under axisymmetric tension on the outer boundary. Numerical results for the circular ring with two diametrally opposed cracks were supplied by Freese and Bowie in a private communication with Underwood [13]. The concept of load relief factor, first used by Neuber [14], was applied by Baratta [15] to determine approximately the stress intensity factors for a multiply cracked circular ring. A weight function technique proposed by Rice [16] was employed by Grandt [17] to obtain stress intensity factors for one and two radially cracked rings.

The finite element technique has become an important numerical method for practical problems in structural mechanics because of its ability to treat very general geometric configurations and loading conditions. Shannon [18], using the constant strain triangular elements, obtained stress intensity factors for a thick-walled cylinder with one or two radial cracks. Because of the large strain gradients in the vicinity of a crack tip, an extremely fine element grid was used near the crack tip. His results were compared with Bowie and Freese's mapping-collocation results in [13]. Many refinements in finite element approach to crack problems have been developed in the past decade. Wilson [19] introduced the embedded singularity. Special crack tip elements were used by Tracey [20], Blackburn [21], Benzley and Beisinger [22]. Using quadratic isoparametric elements, Henshell and Shaw [23] and Barsoum [24] found independently that the inverse square root singularity at a crack tip was obtained by placing the mid-side nodes at quarter points in the vicinity of the crack tip. These elements are implemented in NASTRAN as dummy elements by Hussain et al [25]. A 12-node, isoparametric element is used by Gifford [26], where two special crack tip elements are implemented in his computer program APES [26] for fracture mechanics problems. The concept of shifting the mid-side nodes to quarter points in an 8-node isoparametric element is extended by the authors to 12-node isoparametric elements [27]. The inverse square root singularity of the strain field at the crack tip is obtained by collapsing the quadrilateral elements into triangular elements around the crack tip and placing the two side nodes on each of the straight line segments passing through the tip at 1/9 and 4/9 of the length of the segment from the tip. With these collapsed 12-node traingular elements as singular elements around a crack tip, APES is used in this paper to compute the stress intensity factor for a radially multiple-cracked circular ring. The results are compared with previous results obtained by other investigators.

II. THE 12-NODE QUADRILATERAL ISOPARAMETRIC ELEMENT. A typical 12-node, quadrilateral element, in Cartesian coordinates (x,y) which is mapped to a square in the curvilinear space (ξ,η) with vertices at $(\pm 1, \pm 1)$ is shown in Figure 1. The assumption for displacement components takes the form:

$$u = \sum_{i=1}^{12} N_i(\xi,\eta)u_i$$

$$v = \sum_{i=1}^{12} N_i(\xi,\eta)v_i \tag{1}$$

where u,v are x,y components of displacement of a point whose natural coordinates are ξ,η; u_i,v_i are displacement components of node i and $N_i(\xi,\eta)$ is the shape function given in [28] which can be written as:

$$N_i(\xi,\eta) = \frac{1}{256}(1+\xi\xi_i)(1+\eta\eta_i)[-10+9(\xi^2+\eta^2)][-10+9(\xi_i^2+\eta_i^2)]$$

$$+\frac{81}{256}(1+\xi\xi_i)(1+9\eta\eta_i)(1-\eta^2)(1-\eta_i^2)$$

$$+\frac{81}{256}(1+\eta\eta_i)(1+9\xi\xi_i)(1-\xi^2)(1-\xi_i^2) \quad , \tag{2}$$

for node i whose Cartesian and curvilinear coordinates are (x_i,y_i) and (ξ_i,η_i) respectively. The element is isoparametric, hence the same shape function is used for coordinate transformation,

$$x = \sum_{i=1}^{12} N_i(\xi,\eta)x_i \quad ,$$

$$y = \sum_{i=1}^{12} N_i(\xi,\eta)y_i \quad . \tag{3}$$

The element stiffness matrix is found in the usual way and is given by [24,25]

$$[K] = \int_{-1}^{1}\int_{-1}^{1} [B]^T[D][B] \det |J| d\xi d\eta \tag{4}$$

where [B] is a matrix relating joint displacements to the strain field $\{\varepsilon\}$

$$\{\varepsilon\} = [B] \left\{ \begin{matrix} \cdot \\ \cdot \\ \cdot \\ u_i \\ v_i \\ \cdot \\ \cdot \\ \cdot \end{matrix} \right\}$$

and [D] is the material stiff matrix given by

$$[D] = \frac{E}{1-\nu^2} \begin{bmatrix} 1 & \nu & 0 \\ \nu & 1 & 0 \\ 0 & 0 & (1-\nu)/2 \end{bmatrix} \quad \text{and} \quad [D] = \frac{E}{(1+\nu)(1-2\nu)} \begin{bmatrix} 1-\nu & \nu & 0 \\ \nu & 1-\nu & 0 \\ 0 & 0 & (1-2\nu)/2 \end{bmatrix} \tag{5}$$

for the case of plane stress and plane strain respectively. In (5), E is the modulus of elasticity and ν is Poisson's ratio of the material. The Jacobian matrix [J] is given by

$$[J] = \begin{bmatrix} \frac{\partial x}{\partial \xi} & \frac{\partial y}{\partial \xi} \\ \frac{\partial x}{\partial \eta} & \frac{\partial y}{\partial \eta} \end{bmatrix} = \begin{bmatrix} \cdots \frac{\partial N_i}{\partial \xi} \cdots \\ \cdots \frac{\partial N_i}{\partial \eta} \cdots \end{bmatrix} \begin{bmatrix} \cdot & \cdot \\ \cdot & \cdot \\ x_i & y_i \\ \cdot & \cdot \\ \cdot & \cdot \end{bmatrix} \tag{6}$$

whenever the determinant of [J] is zero, the stresses and strains become singular [23-25]. (This is due to the fact the computation of $\{\varepsilon\}$ requires inverse of [J].)

III. THE CRACK TIP ELEMENT.

The inverse square root singularity of the elastic strain field at a crack tip can be obtained by a simple technique of collapsing the quadrilateral elements into triangular elements around the crack tip as shown in Figure 2. The side nodes 2,3 of the line segment 1-4 of length ℓ are moved to $\ell/9$ and $4\ell/9$ positions measured from the tip, node 1. Similarly, nodes 9 and 8 of the line segment 7-10 of length ℓ (we choose to use isosceles triangles around a crack tip, but any scalene triangles may be used) are moved to $\ell/9$ and $4\ell/9$ from node 10 which coincides with nodes 11, 12 and 1. The line segment 4-7 must be straight and the nodes 5 and 6 must divide the line segment into equal segments. Otherwise numerical results may become unstable [27,29]. From Eq. 3 the Cartesian coordinates of any point (ξ,η), $-1\leq\xi\leq 1$ and $-1\leq\eta\leq 1$, are,

$$x = (\ell/8)(1+\xi)^2[(1-\eta)\cos\beta + (1+\eta)\cos\alpha]$$
$$y = (\ell/8)(1+\xi)^2[(1-\eta)\sin\beta + (1+\eta)\sin\alpha] \tag{7}$$

The Jacobian then becomes,

$$|J| = (\ell/4)^2(1+\xi)^3\sin(\alpha-\beta) \tag{8}$$

This shows the strain is singular at $x = 0$ ($\xi = -1$) along any ray from $x = 0$ since $|J| = 0$ at $\xi = -1$ for all η. It has been shown in [27] that the singularity of the strain field at $r = 0$ is of the order of $(1/\sqrt{r})$ if the nodes 1, 10, 11 and 12 are tied together during deformation, i.e.,

$$u_1 = u_{10} = u_{11} = u_{12} \quad , \quad v_1 = v_{10} = v_{11} = v_{12} \tag{9}$$

If nodes 1, 10, 11 and 12 are not tied together, then the strain singularity is of the order of $(1/r)$, the perfect plastic singularity. This is analogous to the findings in [30] for quadratic, isoparametric elements. However the authors have found that the multipoint constraint has little effect on numerical results for the elastic plane problem of a circular ring with multiple cracks.

IV. DETERMINATION OF STRESS INTENSITY FACTORS.

If displacements of a node near the crack tip obtained from the finite element method are substituted into the well known near crack tip displacement formula [31], the stress intensity factor can be simply computed. There are a number of ways to estimate the stress intensity factors [32] from the nodal displacements. A simple, yet accurate way for a mode I crack is the use of vertical component of the relative displacement of node 14 in reference to node 10, Figure 3.

$$K_I = \left(\frac{2\pi}{\ell}\right)^{\frac{1}{2}} \frac{3E(v_{14}-v_{10})}{(1+\nu)(\kappa+1)} \tag{10}$$

This simple formula gives good results if 1% - 2% of the crack length is used for $(\ell/9)$, the distance between the crack tip and the nearest node.

V. IDEALIZATION OF A RING SECTOR. The internal pressure p applied on the internal bore leads to stress boundary conditions on the boundary of radial cracks. Since the stress intensity factor K_I associated with radial cracks in an internally pressurized cylinder is the same as that produced by uniform axisymmetric tension of equal magnitude applied on the external boundary we consider this external loading in the present finite element method.

A great advantage of using cubic isoparametric elements is that only a few elements are needed to model an elastic structure containing cracks. Let the number of radial cracks be n and let the cracks be of equal depth and equally spaced. The central angle between two adjacent cracks is $2\pi/n$. Let R_1 and R_2 be inner and outer radii of the ring; the diameter ratio R_2/R_1 is denoted by W. The wall thickness of the ring, $t = R_2 - R_1 = R_1$ (W-1), is used to normalize the crack depth a. The crack depth to wall thickness ratio a/t is called the dimensionless crack length, an important parameter used in this analysis. For the axisymmetric problem, the region of interest is confined by $1 \leq R \leq W$ and $0 \leq \theta \leq \theta_0$, where $\theta_0 = \pi/n$ and $R = r/R_1$, Figure 4. The pair (r,θ) are polar coordinates with the center of the cylinder as the pole and the line on which lies the only crack in the region of interest as the polar axis. The region of interest is subdivided into several ring sectors. The maximum number of sectors (NS) used is six and the maximum number of elements (NE) is seventeen. The number of nodes (NN) for 17 elements is 119. The sector containing the crack has seven elements. The remaining sectors have only two elements in each sector. A reduction of two elements and thirteen nodes will result from a reduction of one sector. The sector containing the crack has a central angle <A. The central angle of the remaining sectors may be either <A or <B. Let NA and NB be the number of sectors having central angle <A and <B respectively. The sum of NA and NB is the total number of sectors used for that particular geometry. A table in Figure 4 gives the actual values of NE, NN, NS, NA, NB, <A, <B used in our numerical computation for n = 1 to 40.

Figure 5 gives the actual idealization for a cylinder with two diametrically opposed cracks. The region of interest is a quarter of the cylinder. We have used NE = 17, NN = 119, NS = 6, NA = 1, NB = 5, <A = 7.5°, and <B = 16.5°. The numbering sequence of the seven elements and their nodes varies slightly depending on the dimensionless crack length. For crack length $a/t = \leq 0.6$ the elements are numbered as shown in Figure 5(a). For $a/t > 0.6$, the change is shown in Figure 5(b).

VI. THE COMPUTER PROGRAMS NASTRAN AND APES. The NASTRAN implementation of the 12-node quadrilateral follows that of the 8-node quadrilateral as described in [25]. The dummy user element facility of NASTRAN is used. This requires coding routines to calculate element stiffness matrices and stress recovery computations. Modifications to existing NASTRAN source code are made to provide proper output formats for the element. Stress intensity factors for mode I cracks are calculated using Eq. (10) and using the multipoint constraint, Eq. (9). Three-point (four-point as optional) Gaussian quadrature is normally used to evaluate each partial integrations of the double integral in Eq. (4). All stiffness computations are performed in double precision while stress recovery is performed in single precision.

The finite element computer program APES [26], an acronym for 'Axisymmetric/Planar Elastic Structures', is a special program, using 12-node quadrilateral elements. It automatically generates coordinates of nodes intermediate to element corner nodes for a straight-line element edge. The work-equivalent loads are automatically computed for arbitrarily distributed stresses along any element edge. The result of these features greatly reduces both the effort of input data preparation and the probability of error. Because of the convenience, APES is chosen in the present computation of stress intensity factors for multiple cracks.

For linear elastic fracture mechanics applications, APES has two special crack tip elements for users to choose. The first of these is a circular 'core' element [19] centered on the crack tip in which the leading terms of the elastic singular solution are taken to dominate. The second is simply an enrichment of the 12-node isoparametric element [22]. We believe the collapsed 12-node triangular elements are more convenient to use for crack problems than either of the two special crack tip elements originally used in APES. We modified the program slightly to avoid the use of special crack tip elements. A collapsed triangular element around a crack tip are defined as an ordinary quadrilateral element with four nodes having the same coordinates and four intermediate nodes shifted to $\ell/9$ and $4\ell/9$ positions. The multipoint constraint conditions, Eq. (9) are not yet available in APES. Since the effect in the elastic range is negligibly small, the stress intensity factors are computed by Eq. (10) with nodes 1 to 10 not tied together in the direction of crack. In the direction perpendicular to the crack the conditions $v_1 = v_2 = \dots = v_{10} = 0$ are used.

VII. NUMERICAL RESULTS. The computer program APES, incorporated with collapsed 12-node triangular elements as singular elements around a crack tip, is used to evaluate stress intensity factors for a thick-walled cylinder with one and two radial cracks. Comparing with Bowie and Freese's results the present finite element results are in better agreement than those of Shannon's finite element results. Tables 1 and 2 list values of $K_I/p\sqrt{a}$ for various values of a/t and for W = 1.6 and 2.5. The values in columns (1) and (2) of these tables are either calculated from Tables 1 and 2 of Underwood [13] or from Figure 7-6 of Shannon [33]. The difference between our finite element results and those of Bowie and Freese are, in general, within a two percent range.

The weight function method employed by Grandt also gives accurate results for one and two radial cracks. For W = 2.0, a comparison is listed in Table 3. Values in the column under Grandt are taken from Figure 6 of [17]. The agreement among values obtained from the mapping collocation method, the weight function method and the present finite element method are quite good.

With the confidence thus gained, the computer program APES is used to calculate stress intensity factors of cylinders with many radial cracks. Using the idealization given in Figure 4, values of $K_I/p\sqrt{a}$ are obtained for a/t = 0.1 to 0.6 and for W = 1.5, 2.0 and 2.5. Results are shown in Figures 7 to 9 for n = 1 to 40. There are very few reliable numerical results available in the literature for stress intensity factors of a finite ring with more than two radial cracks. We found only Goldthorpe's results for forty cracks [9] and Baratta's results for thirty-six and forty-eight cracks [15] for W = 2.0. For the purpose of comparison, we obtained $K_I/p\sqrt{a}$ for twenty cracks for a ring of W = 2.0 using 'load relief factors' method [15] and the data for multiple cracks at a circular hole in an infinite solid by Tweed and Rooke [6]. The results and Baratta's results for thirty-six cracks, those of Goldthorpe's forty cracks together with the current finite element results for twenty and forty cracks are plotted in Figure 6. Also in Figure 6 are Bowie and Freese's results for one and two radial cracks. It can be seen that stress intensity factors estimated by Goldthorpe are too low while those by Baratta are too high when compared with results obtained by our finite element method.

In Fig. 7-9 we plotted $K_I/p\sqrt{R_1}$ instead of $K_I/p\sqrt{a}$ versus n . The quantity $K_I/p\sqrt{R_1}$ gives the actual value of K_I for p = 1 and R_1 = 1. An obvious maximum of $K_I/p\sqrt{R_1}$ is seen at n = 2 for a/t > 0.1. The value of $K_I/p\sqrt{R_1}$ decreases rather fast as the number of cracks increases from n = 2. The drop in $K_I/p\sqrt{R_1}$ levels off for n > 20. For a/t > 0.1, it is safe to conclude that the two-crack situation represents the worst case of multiple cracking. The case of three-crack situation is the next worst situation. The single-crack situation is usually worse than the case of four cracks. For a/t = 0.1 the variation in $K_I/p\sqrt{R_1}$ is so small that the differenc between the maximum and the next highest value is within the limit of accuracy of the finite element method. Further numerical results must be obtained to give definite conclusion whether the maximum of $K_I/p\sqrt{R_1}$ always occurs at n = 2 even for very shallow cracks. However it becomes less important whether the worst situation is still the two crack case since the stress intensity factors remain nearly a constant for small values of n for very shallow cracks. In other words in the initial stage of multiple cracking, each crack has little effects on the growth of other cracks.

VIII. CONCLUSIONS.

The use of collapsed 12-node triangular elements as singular crack tip elements has been shown to give excellent results for a thick-walled cylinder with multiple radial cracks. The numerical results shows that the two-crack case is in general the worst situation in multiple cracking. For large crack depths, the stress intensity factor for the worst situation may be as high as one hundred fifty percent of that of the corresponding single crack case. For a very shallow crack, the stress intensity factor of a cracked cylinder will not be significantly affected by the presence of other shallow cracks in a relatively large crack spacing. Hence the single crack case may be used to represent the multiple crack situation when the relative crack depth to crack spacing ratio is small. The effect of interaction of cracks on the stress intensity factor of a cracked cylinder in terms of the dimensionless crack length, the diameter ratio, the relative ratio of crack depth to crack spacing can be studied by the present method using unequal crack depths and unequal spacing of radial cracks.

ACKNOWLEDGEMENT. The authors wish to thank L. N. Gifford, Jr. of Naval Ship Research and Development Center for providing the computer program APES for our use.

REFERENCES.

1. Bowie, O. L., Journal of Mathematics and Physics, Vol. 35, pp. 60-71.

2. Westmann, R. A., Journal of Mathematics and Physics, Vol. 43, pp. 191-198.

3. Kutter, H. K., Int. Journal of Fracture Mechanics, Vol. 6, 1970, pp. 233-247.

4. Tweed, J. and Rooke, D. P., Int. Journal of Engineering Science, Vol. 11, 1973, p. 1185.

5. Tweed, J. and Rooke, D. P., Int. Journal of Engineering Science, Vol. 12, 1974, p. 423.

6. Tweed, J. and Rooke, D. P., Int. Journal of Engineering Science, Vol. 13, 1975, p. 653-661.

7. Winne, D. H., and Wundt, B. M., Trans. ASME, Vol. 80, 1958, pp. 1643-1658.

8. Williams, J. G., and Isherwood, D. P., Journal of Strain Analysis, Vol. 3, 1968, pp. 17-22.

9. Goldthorpe, B. D., "Fatigue and Fracture of Thick Walled Cylinders and Gun Barrels," Case Studies in Fracture Mechanics, edited by T. P. Rich and D. J. Cartright, Army Materials and Mechanics Research Center, Technical Report MS77-5, 1977.

10. Cartright, D. J. and Rooke, D. P., Engineering Fracture Mechanics, Vol. 6, 1974, p. 563.

11. Bowie, O. L., and Neal, D. M., Int. J. Fracture Mechanics, Vol. 6, 1970, p. 199.

12. Bowie, O. L., and Freese, C. E., Journal of Engineering Mechanics, Vol. 4, 1972, p. 315-321.

13. Underwood, J. H., Int. J. Pressure Vessels & Piping, Vol. 3, 1975, p. 229.

14. Neuber, H., Theory of Notch Stresses, AEC TR 4547, 1958.

15. Baratta, F. I., "Stress Intensity Factors for Internal Multiple Cracks in Thick-Walled Cylinders Stressed by Internal Pressure Using Load Relief Factors," Eng. Fracture Mechanics (in Press).

16. Rice, J. R., Int. J. of Solids and Structures, Vol. 8, 1972, pp. 751-758.

17. Grandt, Jr., A. F., "Two Dimensional Stress Intensity Factor Solutions for Radially Cracked Rings," Air Force Materials Lab. AFML-TR-75-121, 1975.

18. Shannon, R. W. E., Int. J. of Pressure Vessels and Piping, Vol. 2, 1974, p. 19.

19. Wilson, W. K., "Combined Mode Fracture Mechanics," Ph.D. Dissertion, University of Pittsburgh, 1969.

20. Tracey, D. M., Engineering Fracture Mechanics, Vol. 3, 1971, pp. 255-265.

21. Blackburn, W. S., "Calculation of Stress Intensity Factors at Crack Tips Using Special Finite Elements, The Mathematics of Finite Elements and Applications," Brunel University, 1973.

22. Benzley, S. E., and Beisinger, A. E., "Chiles - A Finite Element Computer Program that Calculates the Intensities of Linear Elastic Singularities," Sandia Laboratories, Technical Report SLA-73-0894, 1973.

23. Henshell, R. D., and Shaw, K. G., Int. J. Numerical Methods in Eng., Vol. 9, 1975, pp. 495-507.

24. Barsoum, R. S., Int. J. Numerical Methods in Eng., Vol. 10, 1976, pp. 25-37.

25. Hussain, M. A., Lorensen, W. E., and Pflegl, G., NASA TM-X-3428, 1976, p. 419.

26. Gifford, Jr., L. N., Naval Ship Research and Development Center, Report 4799, 1975.

27. Pu, S. L., and Hussain, M. A., "The Collapsed Cubic Isoparametric Element as a Singular Element for Crack Problems," Int. J. Num. Methods in Eng. (in press).

28. Zienkiewiez, O. O., The Finite Element in Engineering Science, McGraw Hill, 1971.

29. Hussain, M. A. and Lorensen, W. E., Proceedings of the 15th Midwest Mechanics Conference, March 1977, p. 40.

30. Barsoum, R. S., Int. J. Numerical Method in Engineering, Vol. 11, 1977.

31. Williams, M. L., Journal of Applied Mechanics, Vol. 24, 1957, pp. 109-114.

32. Pu, S. L., Hussain, M. A., and Lorensen, W. E., "Collapsed 12-Node Triangular Elements As Crack Tip Elements for Elastic Fracture," Watervliet Arsenal Report AR-LCB-TR-77047, 1977.

33. Shannon, R. W. E., "The Application of Linear Elastic Fracture Mechanics to the Internally Pressurized Thick-Walled Cylinder," Ph.D. Dissertation, The Queen's University of Belfast, 1970.

FIGURE CAPTIONS.

Figure 1. Shape Functions and Numbering Sequence for a 12-Node Quadrilateral Element.

Figure 2. A Normalized Square in (ξ,η) Plane Mapped into a Collapsed Triangular Element in (x,y) Plane with the side $\xi = -1$ Degenerated into a Point at the Crack Tip.

Figure 3. Three Collapsed Triangular Elements Surrounding a Mode I Crack Tip.

Figure 4. The Region of Interest for a Ring with n Radial Cracks and the Finite-Element Idealization.

Figure 5. The Idealization of a Ring with Two Radial Cracks and the Numbering Sequence for (a) $a/t \leq 0.6$ and (b) $a/t > 0.6$.

Figure 6. Comparison of Stress Intensity Factors by Different Methods for a Ring with n Radial Cracks.

Figure 7. $K_I/p\sqrt{R_1}$ vs. n for various values of a/t for $W = 2.0$.

Figure 8. $K_I/p\sqrt{R_1}$ vs. n for various values of a/t for $W = 2.5$.

Figure 9. $K_I/p\sqrt{R_1}$ vs. n for various values of a/t for $W = 1.5$.

TABLE 1 - Values of $K_I/p\sqrt{a}$ for pressurized cylinders with one crack

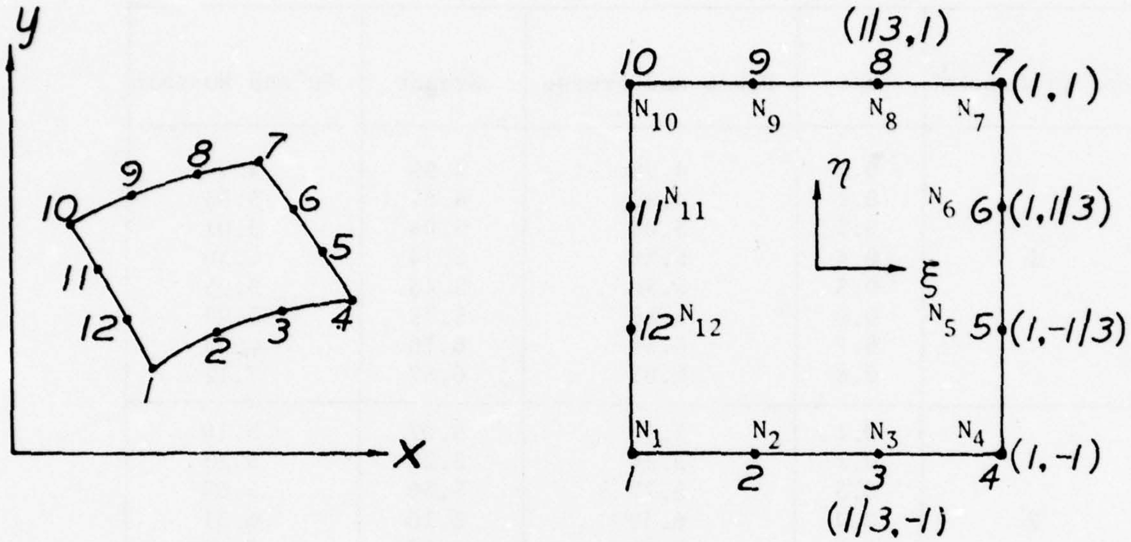
W	a/t	(1) Bowie & Freese	(2) Shannon	(3) Pu and Hussain	(3)/(1)
1.6	0.1	6.39	5.80	6.34	0.992
	0.2	6.58	6.32	6.70	1.018
	0.3	7.10	6.85	7.01	0.987
	0.4	7.82	7.50	7.92	1.013
	0.5	8.60	8.20	8.53	0.992
	0.6	9.38	8.95	9.22	0.983
	0.7	10.16	9.90	10.26	1.010
	0.8	11.01	11.07	11.29	1.025
2.5	0.1	4.16	3.85	4.26	1.024
	0.2	3.97	3.80	4.05	1.020
	0.3	3.92	3.80	3.84	0.980
	0.4	3.97	3.90	3.99	1.005
	0.5	4.06	4.00	4.09	1.007
	0.6	4.25	4.20	4.28	1.007
	0.7	4.58	4.50	4.59	1.002
	0.8	5.10	5.00	5.04	0.988

TABLE 2-Values of $K_I/p\sqrt{a}$ for pressurized cylinders with two cracks

W	a/t	(1) Bowie & Freese	(2) Shannon	(3) Pu and Hussain	(3)/(1)
1.6	0.1	6.39	4.69	6.41	1.003
	0.2	6.91	5.60	7.31	1.032
	0.3	7.75	6.39	7.94	1.025
	0.4	8.86	7.36	9.10	1.027
	0.5	10.29	8.47	10.55	1.025
	0.6	12.05	9.84	11.97	0.993
	0.7	13.88	11.92	14.02	1.010
	0.8	15.90	13.68	16.17	1.017
2.5	0.1	4.30	3.45	4.35	1.012
	0.2	4.35	3.73	4.37	1.005
	0.3	4.59	4.06	4.42	0.963
	0.4	4.96	4.40	4.89	0.986
	0.5	5.48	4.96	5.36	0.978
	0.6	6.10	5.39	6.07	0.995
	0.7	6.90	6.10	6.97	1.010
	0.8	7.94	7.04	8.09	1.019

TABLE 3-Values of $K_I/p\sqrt{a}$ for pressurized cylinders of W = 2.0

No. of Cracks	a/t	Bowie and Freese	Grandt	Pu and Hussain
1	0.1	4.98	4.96	4.99
	0.2	4.92	4.87	5.03
	0.3	5.08	5.04	5.01
	0.4	5.29	5.24	5.36
	0.5	5.56	5.48	5.65
	0.6	5.88	5.79	5.97
	0.7	6.30	6.10	6.35
	0.8	6.93	6.57	7.12
2	0.1	5.03	5.07	5.10
	0.2	5.24	5.23	5.33
	0.3	5.72	5.56	5.60
	0.4	6.30	6.10	6.31
	0.5	7.09	6.82	7.09
	0.6	7.99	7.74	8.08
	0.7	9.11	8.80	9.20
	0.8	10.43	9.92	10.59



$$N_1 = \frac{1}{32}(1-\eta)(1-\xi)[-10+9(\xi^2+\eta^2)]$$

$$N_2 = \frac{9}{32}(1-\eta)(1-\xi^2)(1-3\xi)$$

$$N_3 = \frac{9}{32}(1-\eta)(1-\xi^2)(1+3\xi)$$

$$N_4 = \frac{1}{32}(1-\eta)(1+\xi)[-10+9(\xi^2+\eta^2)]$$

$$N_5 = \frac{9}{32}(1+\xi)(1-\eta^2)(1-3\eta)$$

$$N_6 = \frac{9}{32}(1+\xi)(1-\eta^2)(1+3\eta)$$

$$N_7 = \frac{1}{32}(1+\eta)(1+\xi)[-10+9(\xi^2+\eta^2)]$$

$$N_8 = \frac{9}{32}(1+\eta)(1-\xi^2)(1+3\xi)$$

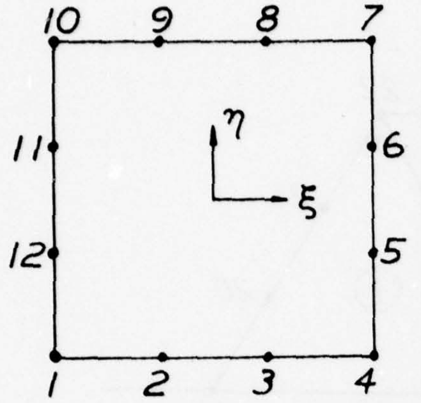
$$N_9 = \frac{9}{32}(1+\eta)(1-\xi^2)(1-3\xi)$$

$$N_{10} = \frac{1}{32}(1+\eta)(1-\xi)[-10+9(\xi^2+\eta^2)]$$

$$N_{11} = \frac{9}{32}(1-\xi)(1-\eta^2)(1+3\eta)$$

$$N_{12} = \frac{9}{32}(1-\xi)(1-\eta^2)(1-3\eta)$$

Figure 1. Shape Functions and Numbering Sequence For a 12-Node Quadrilateral Element.



NODE	x/ℓ	y/ℓ
1	0	0
2	cosβ/9	sinβ/9
3	4cosβ/9	4sinβ/9
4	cosβ	sinβ
5	(2cosβ+cosα)/3	(2sinβ+sinα)/3
6	(cosβ+2cosα)/3	(sinβ+2sinα)/3
7	cosα	sinα
8	4cosα/9	4sinα/9
9	cosα/9	sinα/9
10	0	0
11	0	0
12	0	0

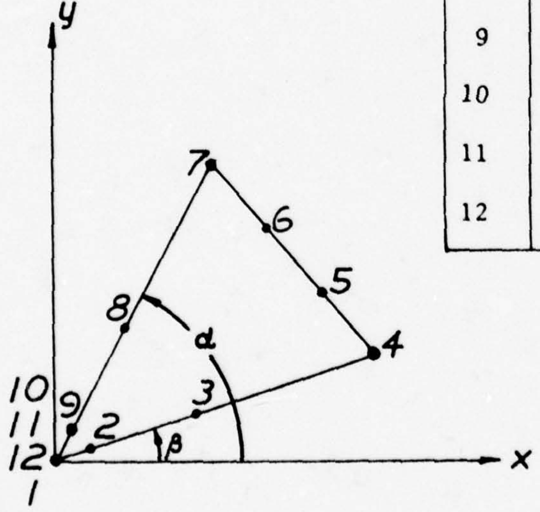


Figure 2. A Normalized Square in (ξ,η) Plane Mapped Into a Collapsed Triangular Element in (x,y) Plane with the side ξ = -1 Degenerated into a Point at the Crack Tip.

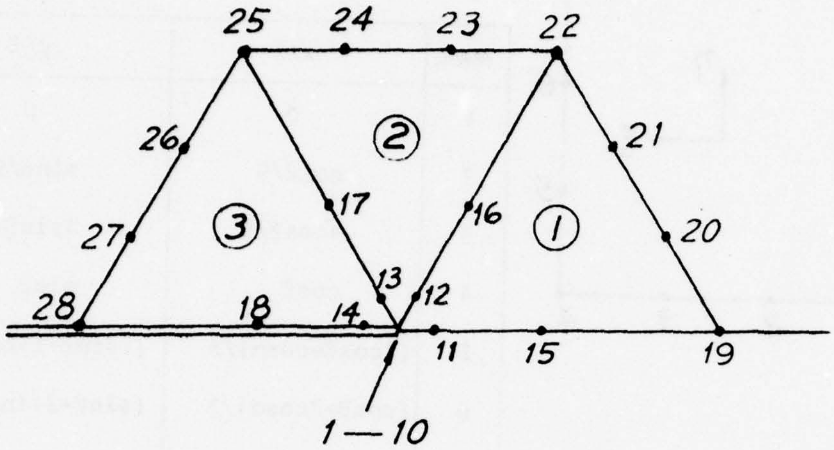
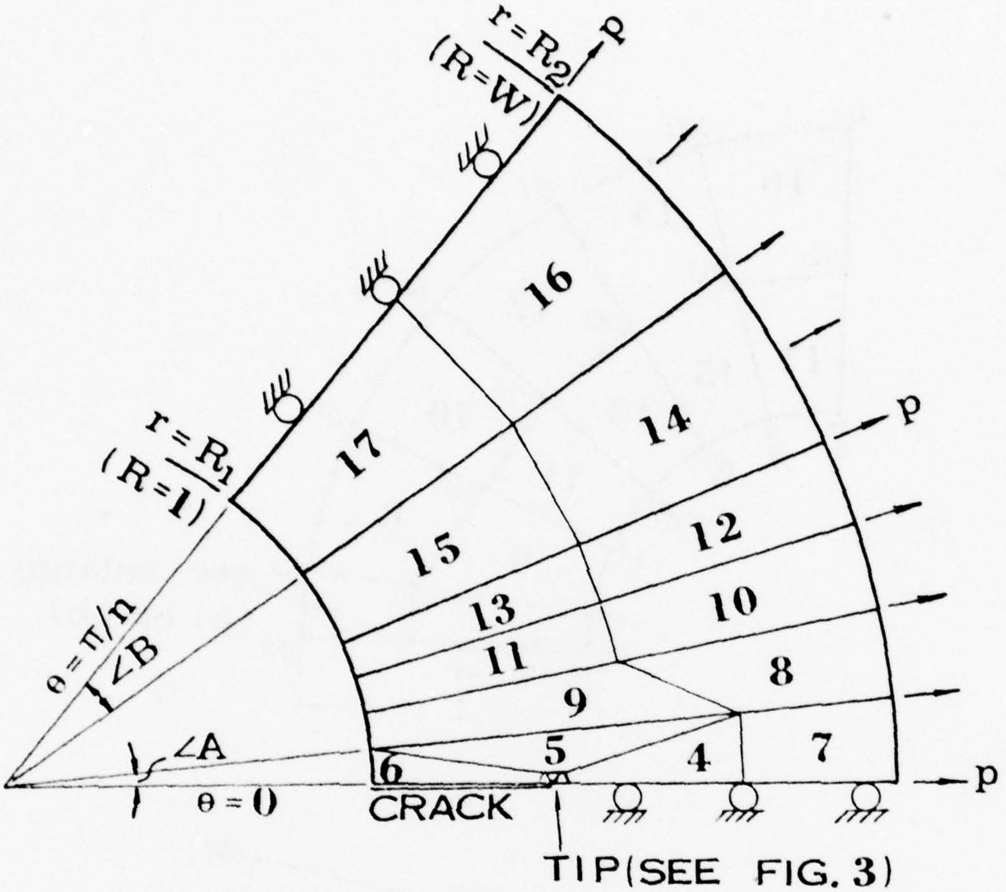


Figure 3. Three Collapsed Triangular Elements Surrounding a Mode I Crack Tip.



n	1	2	3	4	5	6	8	9	10	12	15	18	20	30	40
NE	17									15	13	11		9	
NN	119									106	93	80		67	
NS	6									5	4	3		2	
NA	1	1	2	4	6	4	3	5	6	5	4	2	3	2	1
NB	5	5	4	2	0	2	3	1	0	0	0	1	0	0	1
∠A	7.5		6.0			4.5	3.0								
∠B	34.5	16.5	12.5	10.5	0	6.0	4.5	5.0	0	0	0	4.0	0	0	1.5

Figure 4. The Region of Interest for a Ring with n Radial Cracks and the Finite-Element Idealization.

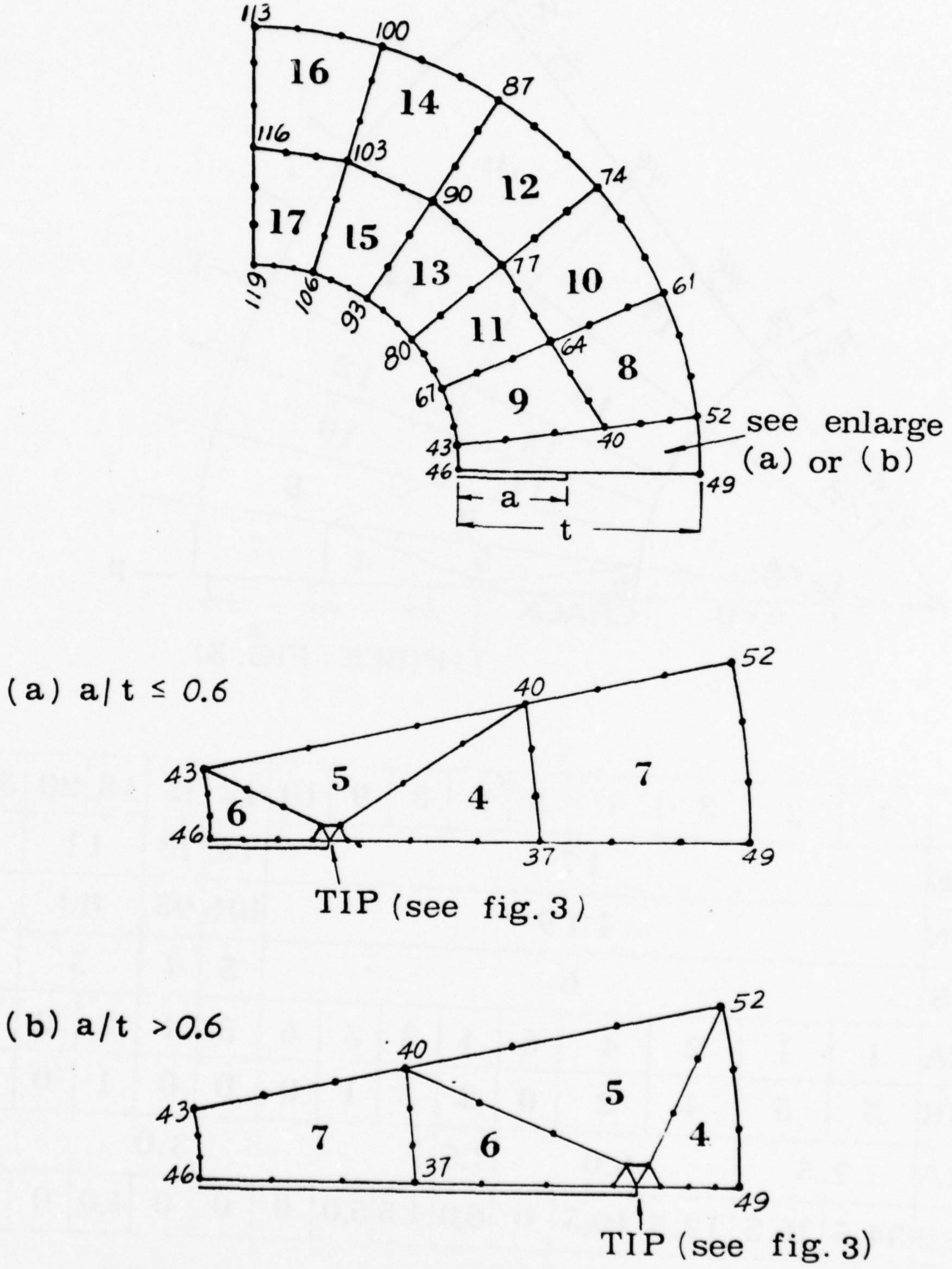


Figure 5. The Idealization of a Ring with Two Radial Cracks and the Numbering Sequence for (a) a/t ≤ 0.6 and (b) a/t > 0.6.

Figure 6. Comparison of Stress Intensity Factors by Different Methods for a Ring with n Radial Cracks.

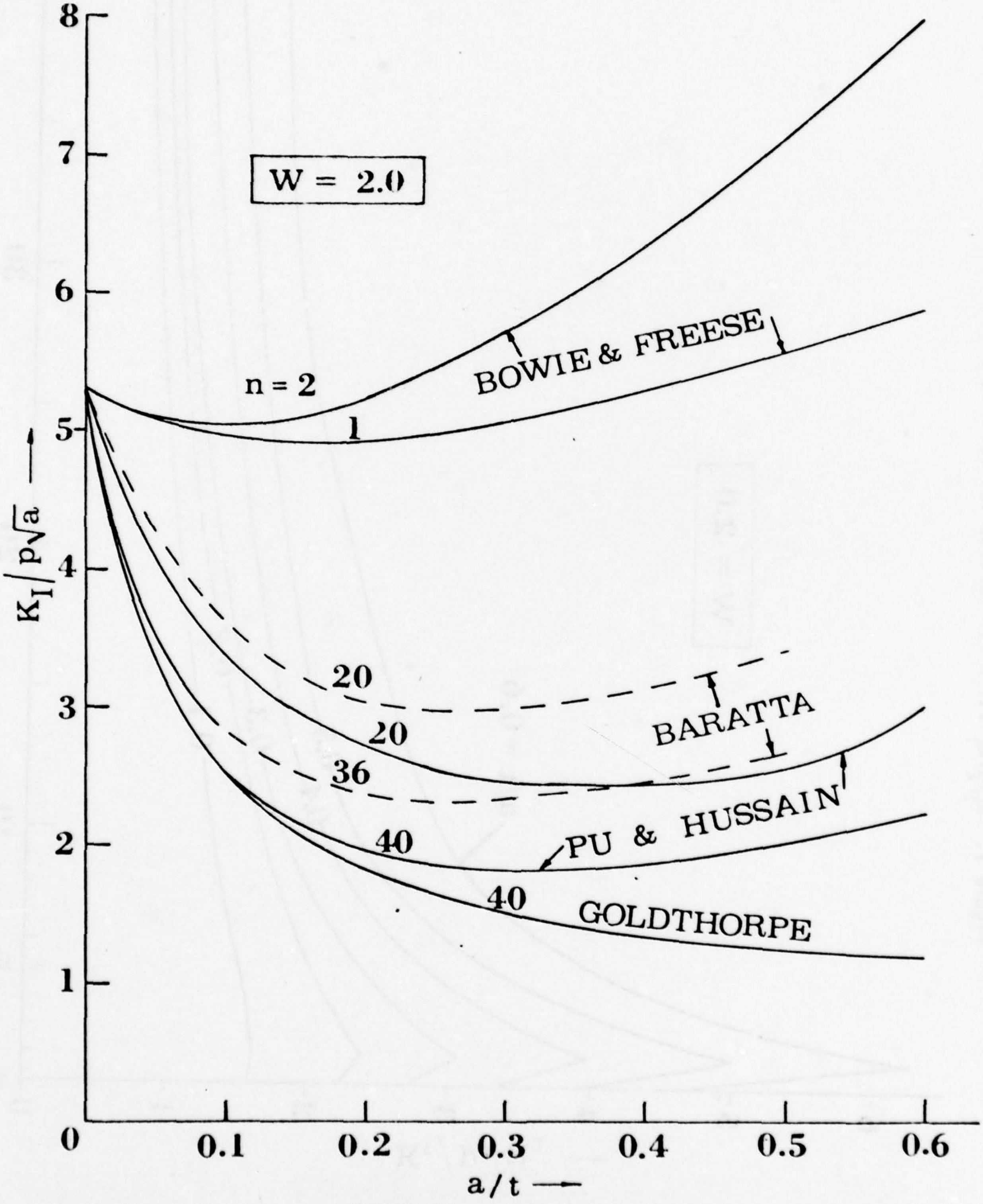


Figure 7. $K_I/p\sqrt{R_1}$ vs. n for various values of a/t for W = 2.0.

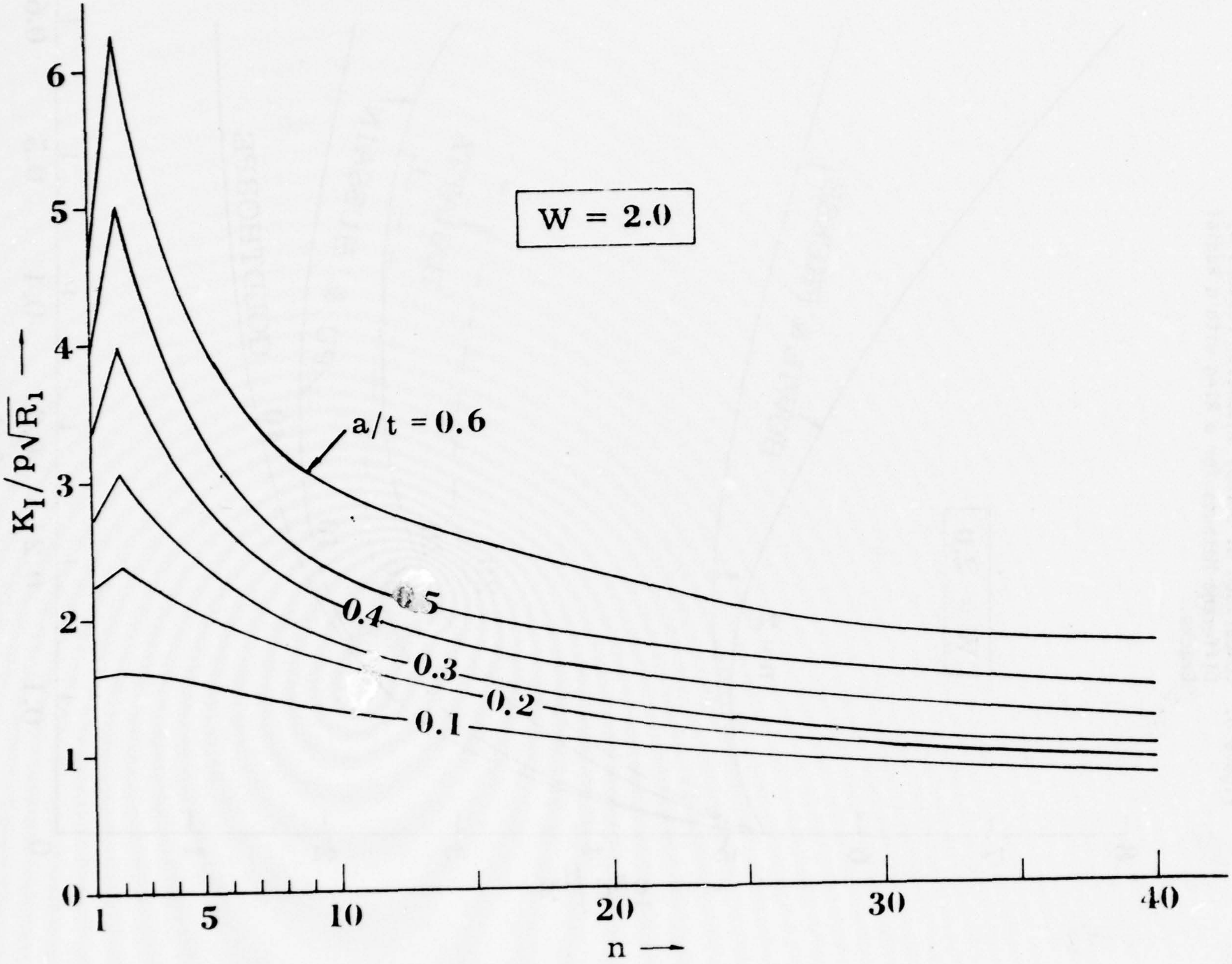


Figure 8. $K_I/p\sqrt{R_1}$ vs. n for various values of a/t for W = 2.5.

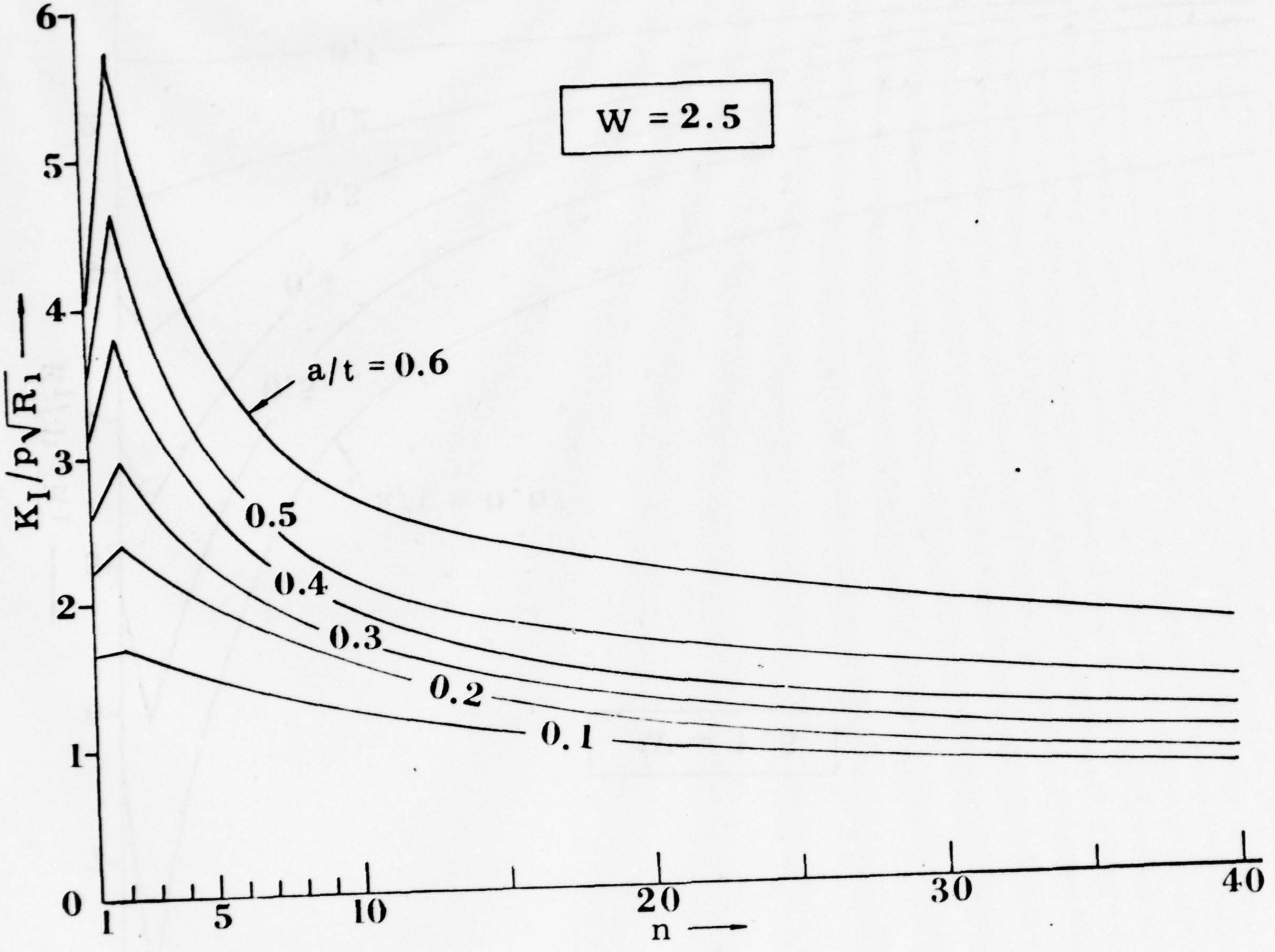
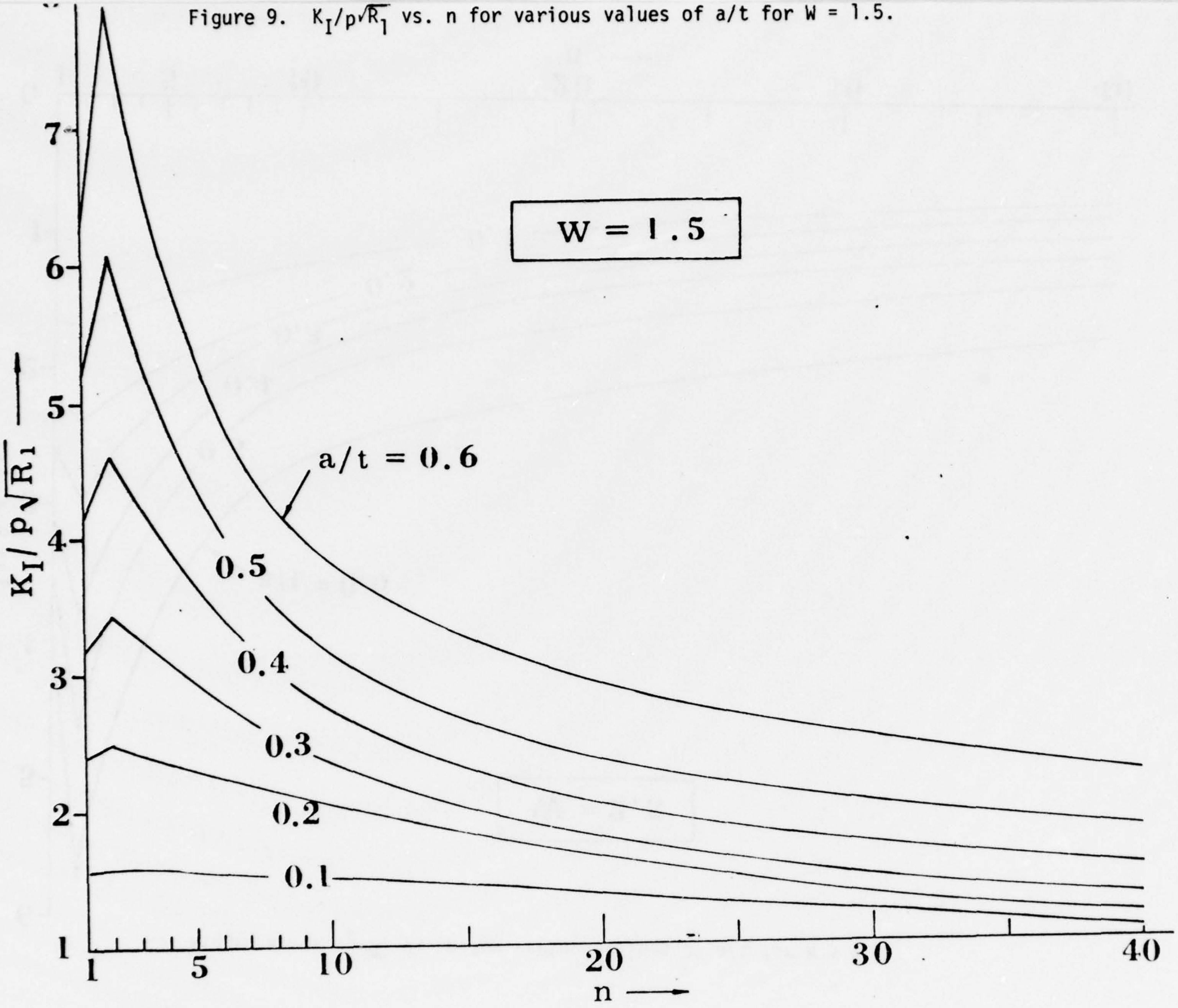


Figure 9. $K_I/p\sqrt{R_1}$ vs. n for various values of a/t for W = 1.5.



TEMPERATURES AND STRESSES DUE TO QUENCHING OF HOLLOW CYLINDERS

John D. Vasilakis
U.S. Army Armament Research and Development Command
Benet Weapons Laboratory, LCWSL
Watervliet Arsenal, Watervliet, NY 12189

ABSTRACT. After forging, gun tube blanks are heated to a high temperature and quenched to near room temperature before tempering to achieve the required material properties. The purpose of the quench is to bypass the knee of the pearlite phase and thus form the desired martensite phase. This program was undertaken to establish cooling curves while the material is being quenched and to compute the thermal and transformation stresses involved.

The temperatures are computed using implicit finite difference schemes. The problem treated is a nonlinear one in radial heat flow. The problem with cylindrical geometry is assumed to be axisymmetric and the coefficients in the equation such as thermal conductivity are treated as functions of temperature. The boundary conditions are written in a general form allowing the use of temperature, convection or heat flux boundary conditions. The nonlinear problem is solved by using two finite difference schemes in tandem. The first computes the temperatures at the n+1/2 time step assuming constant coefficients computed from a previous temperature distribution. This generates a temperature distribution throughout the thickness which is used to compute new coefficients for the second finite difference scheme which calculates the temperature distribution at the n+1 time step. This process is continued until a steady state or some desired level is reached.

At each time step, the program computes the thermal stresses associated with the temperatures. In addition to this, when the temperature reaches a certain level, called martensite start (M_s), the material begins to undergo the martensite transformation. This transformation involves an increase in material volume of about 3%-4%. A simple view of these transformation stresses is taken and the stresses due to this volume change are computed as the temperature cools to below the martensite start temperature throughout the wall thickness.

Results are presented for various boundary conditions including those expected to exist in the quenching facility.

I. INTRODUCTION. There are several techniques available for quenching metals. The object is to develop some desired microstructure in the material. In the Watervliet Arsenal's rotary forge facility, forged cylindrical tubes are heated to an austenitizing temperature of approximately 1550°F and then quenched to form the desired martensitic structure in the material. Both external and internal diameter quenches are utilized. The outside diameter is spray quenched with four water jets in a diametral plane spraying water onto the tube while the tube is rotating. There are several of these planes located along the axis. The bore or inside diameter is quenched by flushing through a nozzle located at one end of the tube.

While the facility was still in the development stage, several tubes (a higher incidence than normal) developed cracks and some of these were interpreted as quench cracks. While the problem was judged to be metallurgical in nature and has been settled, an interest was indicated in understanding the transient temperatures and stresses involved in the quenching problem and this led to the present study.

First, the transient temperature distribution of an axially symmetric hollow cylinder is found. Differences along the axis are assumed to be minor and ignored. The thermal properties are treated as functions of temperature rendering the equation for heat conduction as nonlinear. The finite difference method is used to solve the temperature problem. The Crank-Nicolson equation which is implicitly stable is used.

In the present study, the stresses due to the temperature distribution and the martensite transformation were computed, assuming the problem was elastic and linear. From the computed stresses, it was obvious that some plastic deformation must occur and that an elastic-plastic analysis was required. This work will be performed in a future study.

The stresses due to the transformation are assumed to be strictly due to a change in volume. As the steel transforms from the austenitic structure to a martensitic structure, a volume increase of 3%-4% occurs in the transformed material. This volume increase gives rise to transformation stresses.

The thermal problem and stress problem are treated as being uncoupled. The heat generated from the transformation is small and will have negligible effect on the temperature distribution during the described quenching procedure. The transformation begins when the temperature in the material reaches M_s, the martensite start temperature, and is completed when the material is past the M_f, or martensite finish temperature. Another technique used in quenching is to quench the material to M_s and slowly allow the transformation to take place. In this case, the heat generated during the transformation, might be significant and the coupled problem might need to be considered.

II. PROBLEM STATEMENT. The partial differential equation for temperatures in a hollow cylinder is

$$\frac{1}{r}\frac{\partial}{\partial r}\left(k(u)r\frac{\partial u}{\partial r}\right) = c(u)\rho(u)\frac{\partial u}{\partial t} \qquad (1)$$

where r represents the distance along a radius, u the temperature, and t the time. The thermal conductivity, specific heat and density are represented by k, c and ρ respectively. These properties are assumed to be functions of the temperature. Axial symmetry is assumed and any effects along the axis are ignored.

The initial condition is given by

$$u(r,o) = U_o \qquad (2)$$

where U_o would represent the high austenitizing temperature. The boundary conditions for the problem described would be of the convection type. However, to allow some flexibility in the program, they were written in the following form

$$\begin{aligned} \frac{\partial u}{\partial r} - h_1 u &= -g_1 \quad \text{at } r = a \\ \frac{\partial u}{\partial r} - h_2 u &= -g_2 \quad \text{at } r = b \end{aligned} \qquad (3)$$

where r = a specifies the inside radius and r = b the outside radius of the cylinder. The values h_i and g_i can be varied at either surface so that various boundary conditions can be specified. For example, if $g_1 = 0$ at r = a, then h_1 is the Nusselt number and convection boundary condition is indicated at $r = a$. If h_2 and g_2 are very large but the ratio $g_2/h_2 = U_b$ then the temperature U_b is specified at $r = b$.

Since the thermal properties of the material must be considered as functions of temperature, the partial differential equation (1) is nonlinear and numerical techniques are needed to solve the problem. An implicit scheme based on the Crank-Nicolson equation was used in writing the finite difference scheme for the temperatures.

III. FINITE DIFFERENCE EQUATIONS*. The Crank-Nicolson representation of Eq. (1) is

$$\frac{1}{(a+(i-\frac{1}{2})\Delta r)\Delta r}\left\{\frac{a+i\Delta r}{2}k_{i+\frac{1}{2},n+\frac{1}{2}}\delta_r(u_{i+\frac{1}{2},n+1}+u_{i+\frac{1}{2},n}) - \right.$$

$$\left.\frac{a+(i-1)\Delta r}{2}k_{i-\frac{1}{2},n+\frac{1}{2}}\delta_r(u_{i-\frac{1}{2},n+1}+u_{i-\frac{1}{2},n})\right\} = c_{i,n+\frac{1}{2}}\rho_{i,n+\frac{1}{2}}\frac{u_{i,n+1}-u_{i,n}}{\Delta t} \tag{4}$$

where i is the ith node
n is the time step
Δt is the time increment
Δr is the space increment

and

$$\delta_r u_{i+\frac{1}{2},n} = \frac{u_{i+1,n}-u_{i,n}}{\Delta r}$$

$$k_{i+\frac{1}{2},n+\frac{1}{2}} = k\left(\frac{u_{i+1,n+\frac{1}{2}}+u_{i,n+\frac{1}{2}}}{2}\right) \tag{5}$$

The finite difference equation (4) is written about the point r_i, $t_{n+\frac{1}{2}}$. If the temperature and its spacial derivatives can be written without requiring their values at $n+\frac{1}{2}$, then the equations become linear [1]. This is accomplished by arithmetic averaging the finite difference analogues at the points r_i, t_n and r_i, t_{n+1}, and the resulting analogue is the average of the forward and backward analogues. To solve these equations for the temperatures, it is required

*Reference 1 was found to be a very useful book on the subject.

to know the properties at the n+½ time step. This will be shown later.

In writing the boundary conditions as Equations (3), it is necessary to locate the nodes as shown in Figure 1. There are no nodes located on the boundary about which finite difference equations are written. The boundary conditions are used to eliminate the 0th and R+1st nodes from the equations. The temperatures on the boundary are found from extrapolation or through the use of the boundary conditions after the spatial temperature distribution is found at that time step.

The values of the thermophysical properties at the ½ time step can be found through various projection methods [1]. The one chosen is the centered Taylor series projection. A set of equations similar to Equation (4) are written between the n and n+½ time step. Thus the values of the properties would thus be required at the n+½ time level. The technique allows the computation to take place using properties evaluated at the known nth time level. Under those conditions, the equations are linear, the properties known and the temperatures at the n+½ time step can be found. Knowing this, new property values can be found and the equations solved for the temperatures at the n+1 time step.

An alternate technique which still arrives at the equivalent of Equation (4) was used for the centered Taylor series projection. Equation (1) is rewritten in following form.

$$k(u)\,\frac{\partial^2 u}{\partial r^2} + \frac{\partial k(u)}{\partial u}\,\left(\frac{\partial u}{\partial r}\right)^2 + \frac{k(u)}{r}\,\frac{\partial u}{\partial r} = \rho(u)c(u)\,\frac{\partial u}{\partial t} \qquad (6)$$

Using the Crank-Nicolson finite difference analogue about the ith node and n+½ time step, Equation (6) becomes, after some rearranging,

$$\frac{1}{2(\Delta r)^2}\,[k(u_{i,n+\frac{1}{2}}) + k'(u_{i,n+\frac{1}{2}})[\frac{u_{i+1,n+\frac{1}{2}} - u_{i-1,n+\frac{1}{2}}}{2\Delta r}]\,\frac{\Delta r}{2}](u_{i+1,n+1} + u_{i+1,n})$$

$$+ \frac{1}{2(\Delta r)^2}\,[k(u_{i,n+\frac{1}{2}}) - k'(u_{i,n+\frac{1}{2}})[\frac{u_{i+1,n+\frac{1}{2}} - u_{i-1,n+\frac{1}{2}}}{2\Delta r}]\,\frac{\Delta r}{2}](u_{i-1,n+1} + u_{i-1,n})$$

(7)

$$- \frac{1}{(\Delta r)^2} k(u_{i,n+\frac{1}{2}})(u_{i,n+1} + u_{i,n}) +$$

$$+ \frac{k(u_{i,n+\frac{1}{2}})}{2(a+(i-\frac{1}{2})\Delta r)} (\frac{u_{i+1,n+1} + u_{i+1,n} - u_{i-1,n+1} - u_{i-1,n}}{2\Delta r})$$

$$= c(u_{i,n+\frac{1}{2}})\rho(u_{i,n+\frac{1}{2}}) \frac{u_{i,n+1} - u_{i,n}}{\Delta t} \qquad (7)$$

However, the coefficients of the first 2 terms can be viewed as truncated Taylor series for $k(u_{i+\frac{1}{2},n+\frac{1}{2}})$

$$k(u_{i+\frac{1}{2},n+\frac{1}{2}}) = k(u_{i,n+\frac{1}{2}}) + \frac{\Delta r}{2} k'(u_{i,n+\frac{1}{2}})(\frac{u_{i+1,n+\frac{1}{2}} - u_{i-1,n+\frac{1}{2}}}{2\Delta r}) \qquad (8)$$

Rewritting Eq. (7)

$$\frac{1}{2(\Delta r)^2} [k(u_{i+\frac{1}{2},n+\frac{1}{2}})](u_{i+1,n+1} + u_{i+1,n}) +$$

$$+ \frac{1}{2(\Delta r)^2} [k(u_{i-\frac{1}{2},n+\frac{1}{2}})](u_{i-1,n+1} + u_{i-1,n}) -$$

$$- \frac{1}{2(\Delta r)^2} [k(u_{i+\frac{1}{2},n+\frac{1}{2}}) + k(u_{i-\frac{1}{2},n+\frac{1}{2}})](u_{i,n+1} + u_{i,n}) + \qquad (9)$$

$$+ \frac{k(u_{i+\frac{1}{2},n+\frac{1}{2}})}{4(a+(i-\frac{1}{2})\Delta r)} (\frac{u_{i+1,n+1} + u_{i+1,n}}{2\Delta r}) - \frac{k(u_{i-\frac{1}{2},n+\frac{1}{2}})}{4(a+(i-\frac{1}{2})\Delta r)} (\frac{u_{i-1,n+1} + u_{i-1,n}}{2\Delta r}) -$$

$$- [c(u_{i+\frac{1}{2},n+\frac{1}{2}}) + c(u_{i-\frac{1}{2},n+\frac{1}{2}})][\rho(u_{i+\frac{1}{2},n+\frac{1}{2}}) + \rho(u_{i-\frac{1}{2},n+\frac{1}{2}})]$$

$$[\frac{u_{i,n+1} - u_{i,n}}{\Delta t}] = 0$$

where

$$k(u_{i,n+\frac{1}{2}}) = \tfrac{1}{2}[k(u_{i+\frac{1}{2},n+\frac{1}{2}}) + k(u_{i-\frac{1}{2},n+\frac{1}{2}})] \tag{10}$$

Equation (9) is now rewritten for evaluation of temperatures at the n+½ time interval using for the coefficients their known values at time step n

$$\frac{1}{2(\Delta r)^2}\left[1 + \frac{\Delta r}{4(a+(i-\frac{1}{2})\Delta r)}\right] k(u_{i+\frac{1}{2},n})(u_{i+1,n+\frac{1}{2}} + u_{i+1,n}) + \frac{1}{2(\Delta r)^2}$$

$$\left[1 - \frac{\Delta r}{4(a+(i-\frac{1}{2})\Delta r)}\right] k(u_{i-\frac{1}{2},n})(u_{i-1,n+\frac{1}{2}} + u_{i-1,n}) - \frac{1}{2(\Delta r)^2} *$$

$$[k(u_{i+\frac{1}{2},n}) + k(u_{i-\frac{1}{2},n})](u_{i,n+\frac{1}{2}} + u_{i,n}) - \tfrac{1}{4}[c(u_{i+\frac{1}{2},n}) + c(u_{i-\frac{1}{2},n})] *$$

$$[\rho(u_{i+\frac{1}{2},n}) + \rho(u_{i-\frac{1}{2},n})] \frac{u_{i,n+\frac{1}{2}} - u_{i,n}}{\Delta t} = 0 \tag{11}$$

This yields the temperature distribution at n+½ using coefficients evaluated from the temperature distribution at time n. The temperatures from Eq. (11) are then used to evaluate the coefficients for use in Eq. (4).

The finite difference equations were tridiagonal and were solved using the Thomas algorithm [1]. Briefly, the form of this algorithm is

$$\begin{array}{ll} a_i u_{i-1} + b_i u_i + c_i u_{i+1} = d_i & 1 \leq i \leq R \\ a_1 = 0, \quad c_R = 0 & \end{array} \tag{12}$$

where the terms on the left hand side are at the n+1 time step and on the right hand side at n time step. For all i, the quantities

$$\beta_i = b_i - \frac{a_i c_{i-1}}{\beta_{i-1}} \;, \quad \gamma_i = \frac{d_i - a_i \gamma_{i-1}}{\beta_i} \tag{13}$$

as computed and then back substitution is used to find the temperatures from

$$u_R = \gamma_R$$
$$u_i = \gamma_i - \frac{c_i u_{i+1}}{\beta_i} \qquad (14)$$

Computation times are rapid.

IV. THERMAL AND TRANSFORMATION STRESSES. The quenching process gives rise to thermal stresses due to the large thermal gradients that exist. Areas near the boundary are cooler than interior points. The boundary would like to contract but is partially prevented from doing so because of the interior, hence tensile stresses are set up near the boundaries while the interior is in compression. The thermal stresses in an axially symmetric hollow cylinder subject to a non-uniform temperature distribution are given by [2].

$$\sigma_r = \frac{E\alpha}{r^2}\left[\frac{r^2-a^2}{b^2-a^2}\int_a^b \rho u(\rho)d\rho - \int_a^r \rho u(\rho)d\rho\right]$$
$$\sigma_\theta = \frac{E\alpha}{r^2}\left[\frac{r^2+a^2}{b^2-a^2}\int_a^b \rho u(\rho)d\rho - \int_a^r \rho u(\rho)d\rho - r^2u(r)\right] \qquad (15)$$

where σ_r.....the radial stresses
σ_θ.....the tangential stresses
E......Young's Modulus
α......thermal expansion coefficient
$u(\rho)$...radial temperature distribution

The stresses due to the transformation are found using similar equations since these stresses are due mainly to a volume increase in the transformed material. The difference between the two calculations is that the transformation does not occur across the thickness simultaneoulsy but progresses across based on the temperature in the cross sections. No transformation stresses exist when the temperatures are all above that temperature (M_s) when the transformation begins or below that temperature (M_f) for which the transformation ends. Between these two temperatures, a linear change in volume is assumed. The change in volume, about 4% if the transformation is complete, is assumed to be isotropic so that it translates to one-third of the volumo change for a linear change during the transformation. Stresses are computed in a manner similar to thermal stresses.

$$\sigma_r = \frac{E}{r^2}\left[\frac{r^2-a^2}{b^2-a^2}\int_a^b \frac{\Delta\ell}{\ell}\,\rho d\rho - \int_a^r \frac{\Delta\ell}{\ell}\,\rho d\rho\right] \tag{16}$$

$$\sigma_\theta = \frac{E}{r^2}\left[\frac{r^2+a^2}{b^2-a^2}\int_a^b \frac{\Delta\ell}{\ell}\,\rho d\rho + \int_a^r \frac{\Delta\ell}{\ell}\,\rho d\rho - r^2\,\frac{\Delta\ell}{\ell}\right]$$

where

$$\begin{array}{lll} \text{I.} & \frac{\Delta\ell}{\ell} = 0 & \text{if } u(r) \geq M_s \\ \text{II.} & \frac{\Delta\ell}{\ell} = \left(\frac{\Delta\ell}{\ell}\right)^* \frac{1}{M_s - M_f}\,(M_s - u(r)) & M_s \geq u(r) \geq M_f \\ \text{III.} & \frac{\Delta\ell}{\ell} = \left(\frac{\Delta\ell}{\ell}\right)^* & M_f \geq u(r) \end{array} \tag{17}$$

and $\left(\frac{\Delta\ell}{\ell}\right)^*$ is the linear expansion during a transformation.

The M_s temperature was taken to be 350°F and the M_f temperature 150°F for the computations. Figure 2 shows some temperature distributions which can arise. In the upper figure, no transformation has taken place, hence the transformation stresses are zero. In the lower figure, the transformation is occurring from both the inside and outside radius. In sections indicated by I, corresponding to Equations (17), the transformation has not begun, in sections II, the transformation is progressing and in sections III, the transformation is complete.

As stated above, for the present study the thermal and transformation stresses were assumed to be elastic. In the following, the results indicate that stresses are too large for this assumption to be valid. References [3] and [4] treat similar problems using elastic-plastic analysis.

V. RESULTS AND DISCUSSION. Figures 3 through 7 show some resulting temperature distributions under various conditions.

Figure 3 shows the temperature distribution across the wall thickness for various times. The boundary conditions are convective and h_2 represents a value near that suggested by the manufacturer of the quenching facility for the coefficient on the outside diameter. On the inside diameter, the value of h_1 was said to be lower than that of h_2. The radius is in inches and the temperatures are in °F. The ambient temperature was assumed zero. Figure 4 shows the effect of variations in the convection coefficient on the inside diameter. The results are shown for only one time step. Since the temperature at that time does not change much under the different boundary conditions, small differences in the convection parameter on the inside diameter will have little effect on the transformation.

Figure 5 shows the effect when the thermal conductivity is allowed to vary with temperature. For the same time step, three curves are shown. The conductivity is allowed to be an increasing, decreasing or constant function of temperature. Finding real data to use in the program is difficult. Figure 6 shows the temperature distribution for a bilinear thermal conductivity curve based on one for 4130 steel. These properties are usually determined experimentally under equilibrium conditions. Since the structure of the material is changing under rapid cooling and since equilibrium does not exist, the properties which should be used are those determined under the same conditions as the quench. This can be described best by looking at the specific heat. Again for 4130 steel, a spike increase in the value of the specific heat occurs between 1200°F and 1500°F. This occurs during heating and is due to the austenitizing of the material. Figure 7 shows the temperature distribution throughout the tube wall, allowing the specific heat to be a function of the temperature but ignoring the spike mentioned above. Under the quench conditions, the spike would occur during the martensite transformation and be of different shape.

Figures 8 and 9 show a sample of the radial and circumferential or hoop stress, respectively, for a specific time. They are taken from one of the previously cited cases. The insert shows qualitatively the temperature distribution at the time indicating that the transformation has begun at both the inside and outside radius. Each figure shows the thermal, transformation and sum of the stresses. Large compressive hoop stresses indicate the strong possibility of plastic deformation. Since the transformation occurs at the lower temperatures, the thermal gradients are smaller and the thermal stresses lower than their values earlier in the quenching cycle. The thermal stresses, however, can be a significant part of the total stresses, especially early in the transformation, and should not be neglected in an elastic-plastic analysis.

ACKNOWLEDGEMENT. The author wishes to acknowledge the assistance of Richard Haggerty and Royce Soanes in the computational phase of this work.

REFERENCES.

1. von Rosenberg, D. U., Methods for the Numerical Solution of Differential Equations,

2. Boley, B. A. and Weiner, J. H., Theory of Thermal Stresses, John Wiley and Sons, Inc., 1960.

3. Weiner, J. H., and Huddleston, J. V., Transient and Residual Stresses in Heat-Treated Cylinders, Journal of Applied Mechanics, March, 1959.

4. Landau, H. G., and Zwicky, E. E. Jr., Transient and Residual Thermal Stresses in an Elastic-Plastic Cylinder, Journal of Applied Mechanics, Sept., 1960.

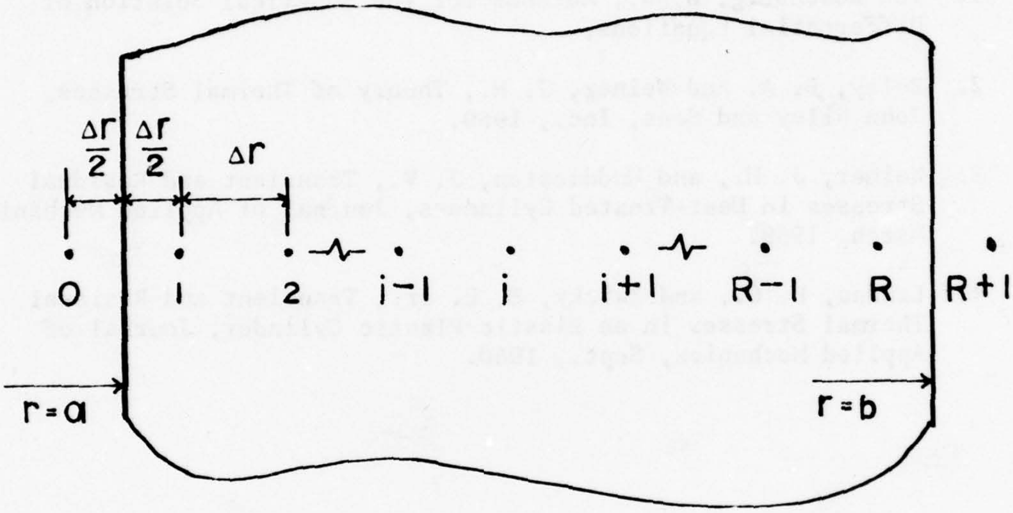


FIGURE 1. NODE PLACEMENT WITH

POINTS SHIFTED FROM BOUNDARIES

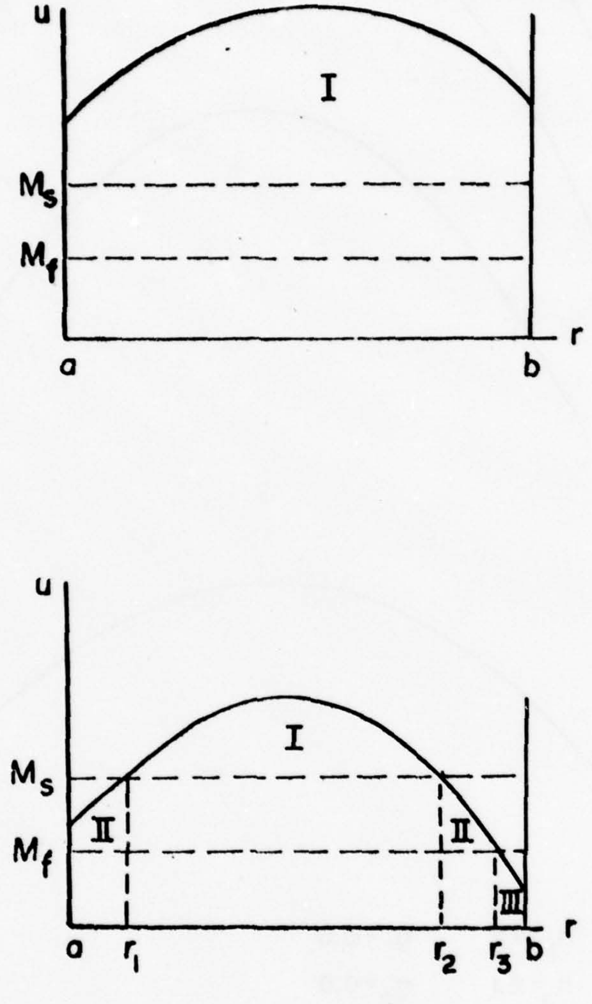


FIGURE 2. TYPICAL TEMPERATURE DISTRIBUTIONS NEAR TRANSFORMATION TEMPERATURE

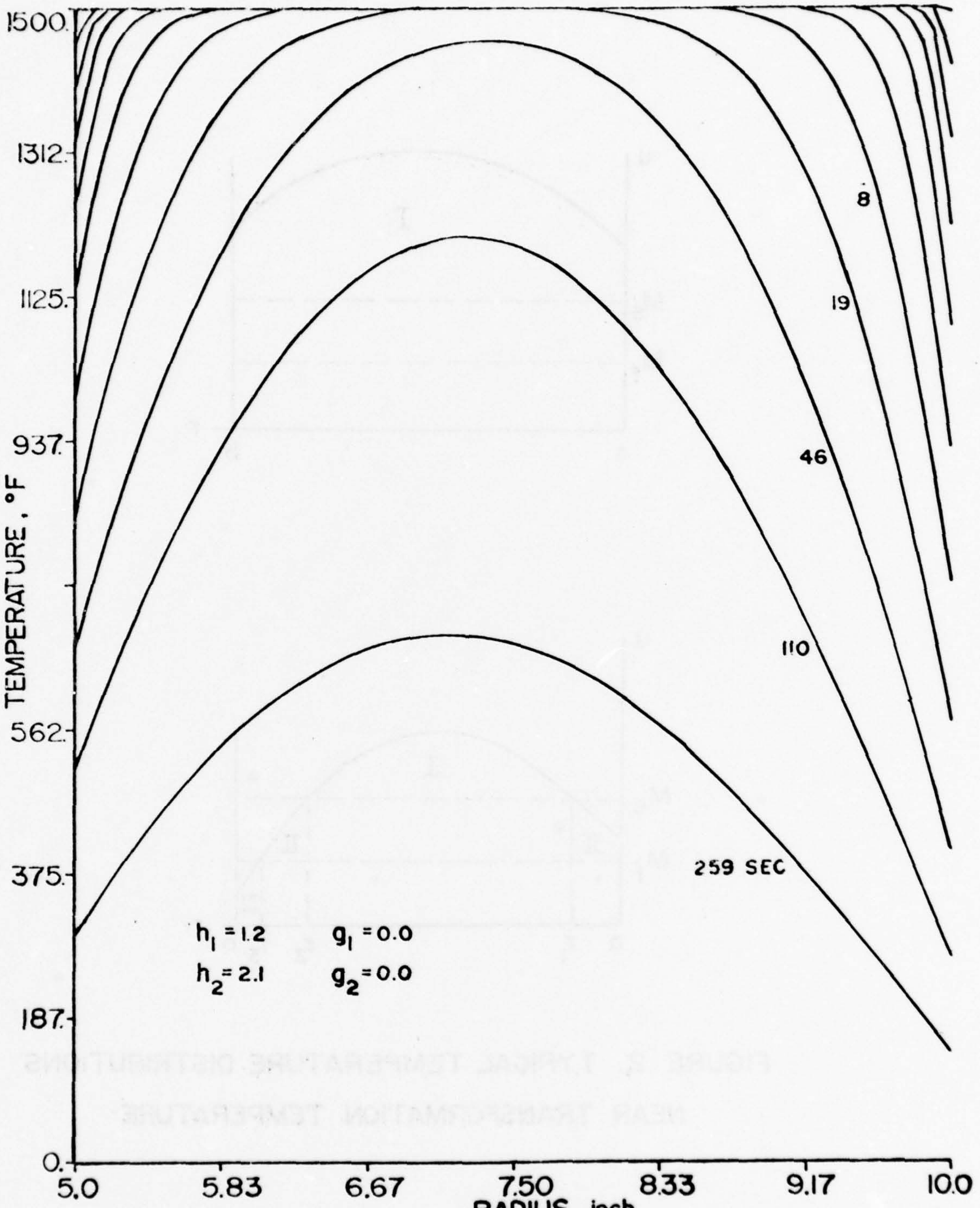


FIGURE 3. TEMPERATURE VS RADIUS FOR VARIOUS TIMES

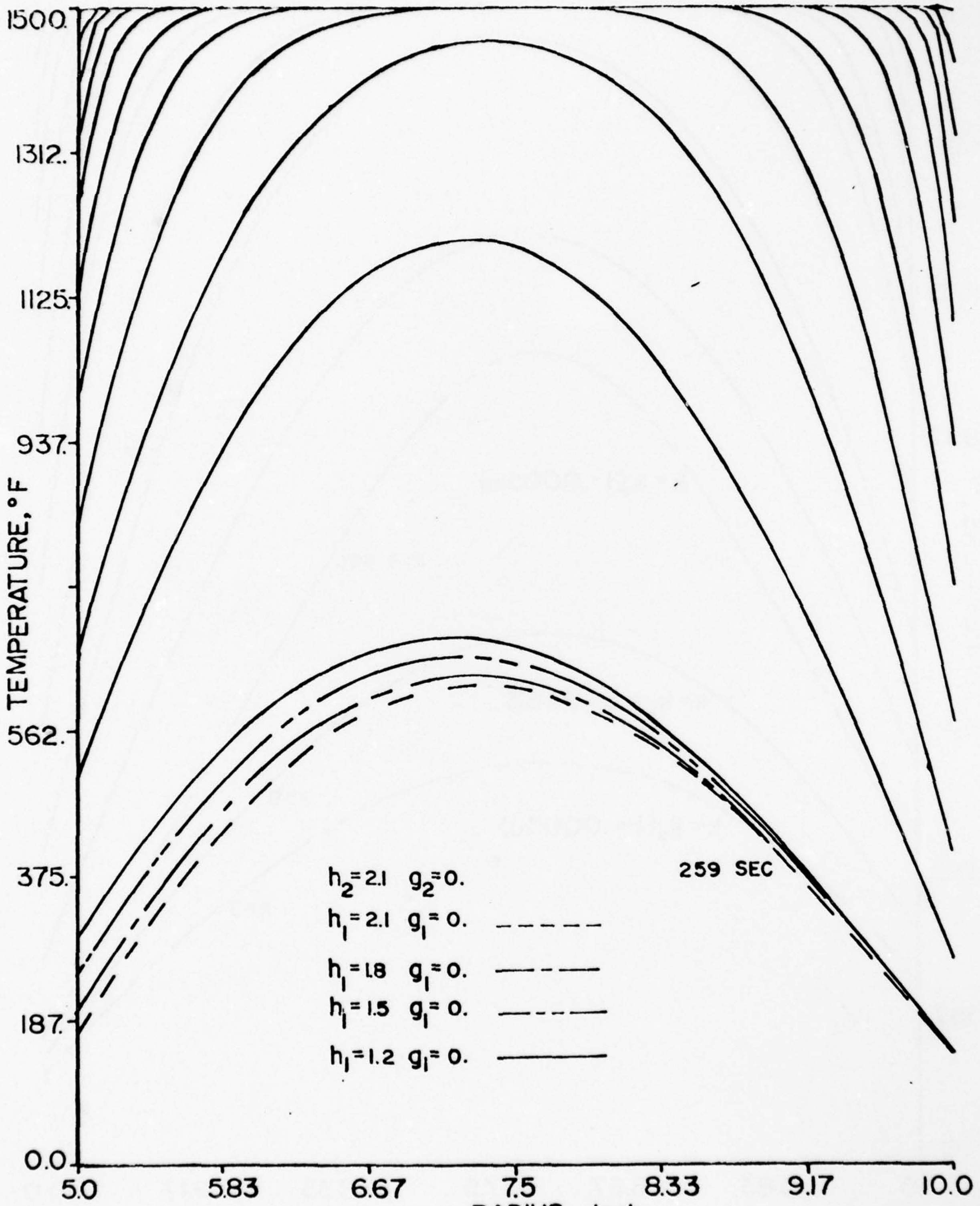


FIGURE 4. VARIATION OF BORE CONVECTION COEFFICIENT

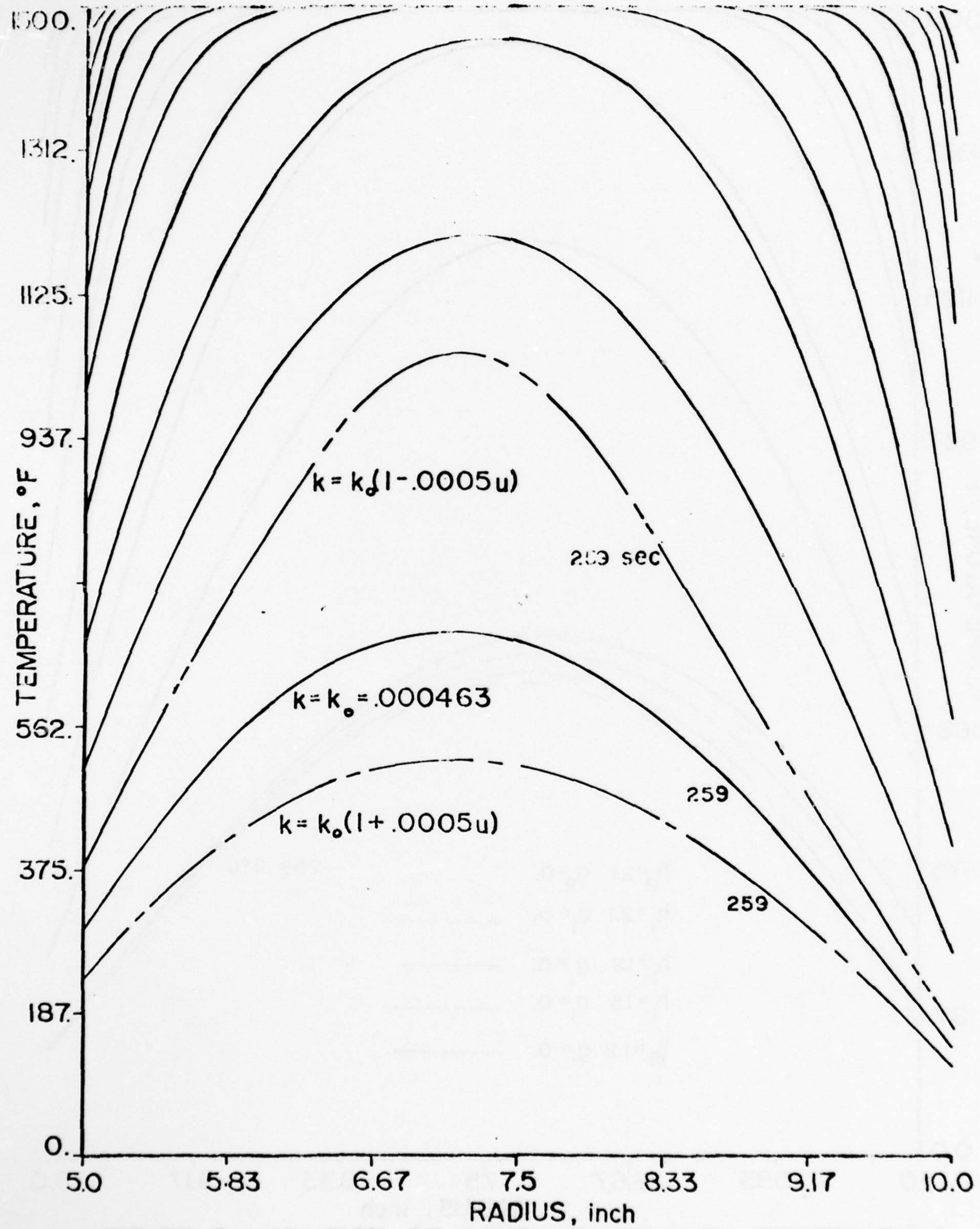


FIGURE 5. VARIATION OF CONDUCTIVITY WITH TEMPERATURE

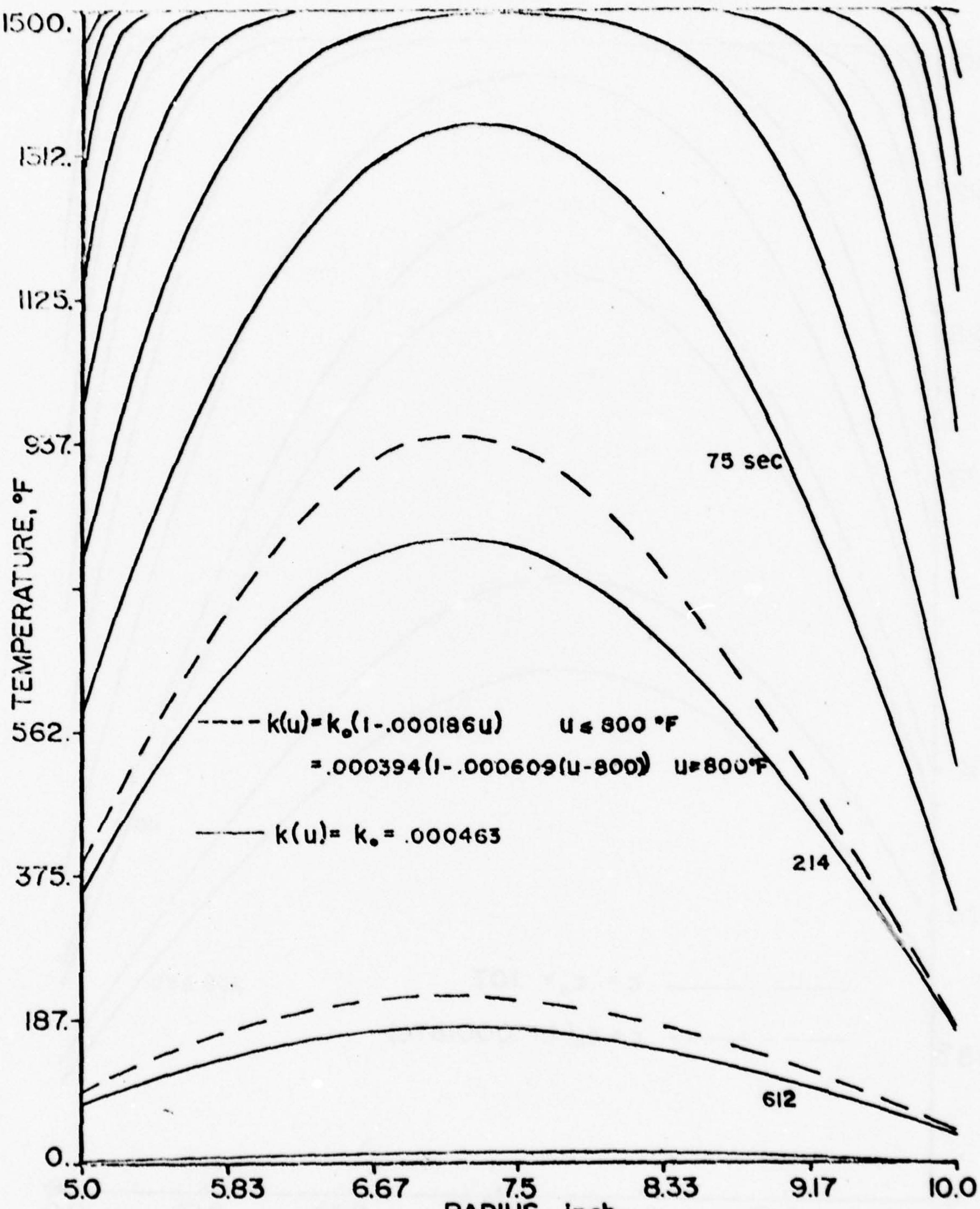


FIGURE 6. VARIATION OF CONDUCTIVITY WITH TEMPERATURE

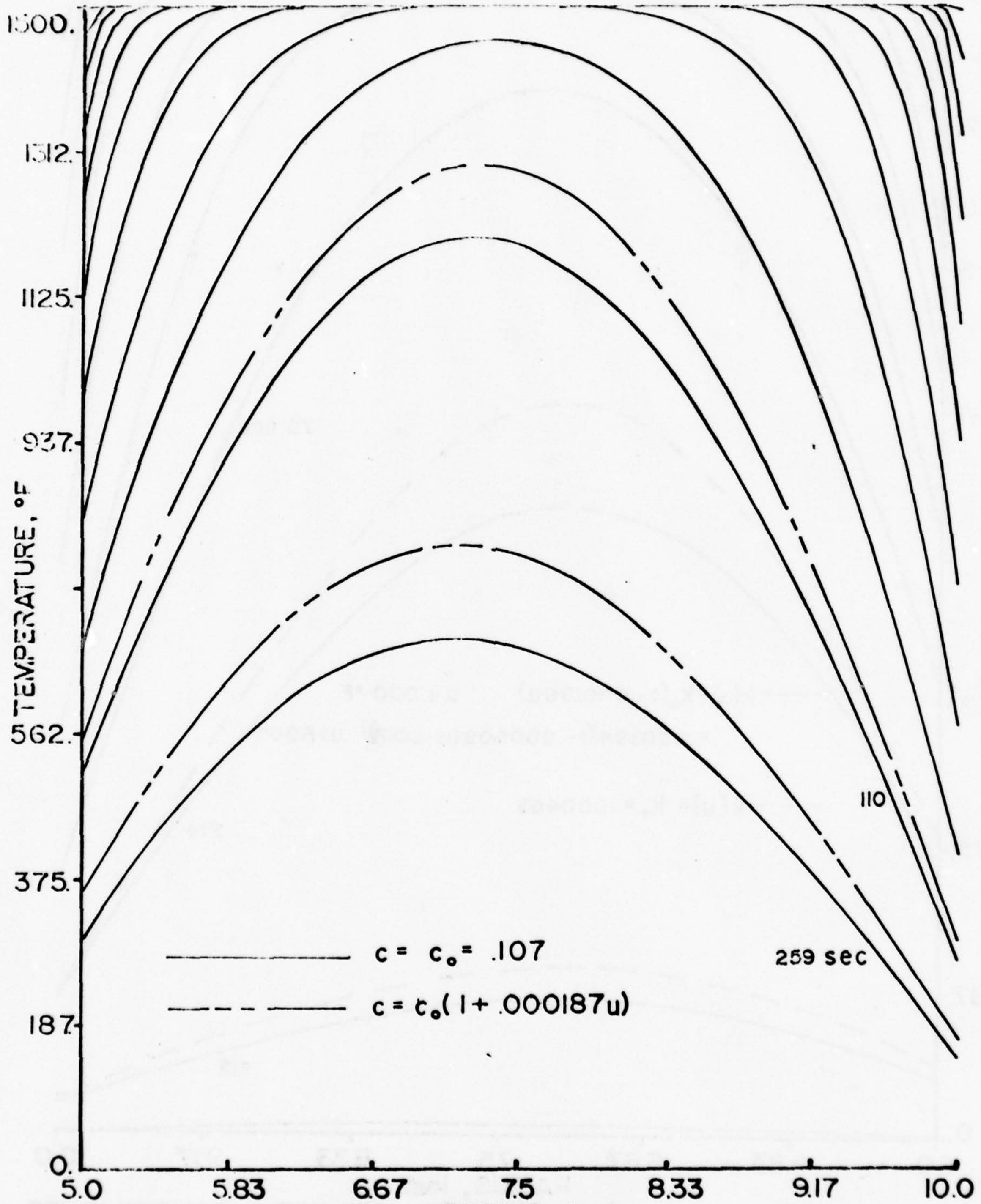


FIGURE 7. VARIATION OF SPECIFIC HEAT WITH TEMPERATURE

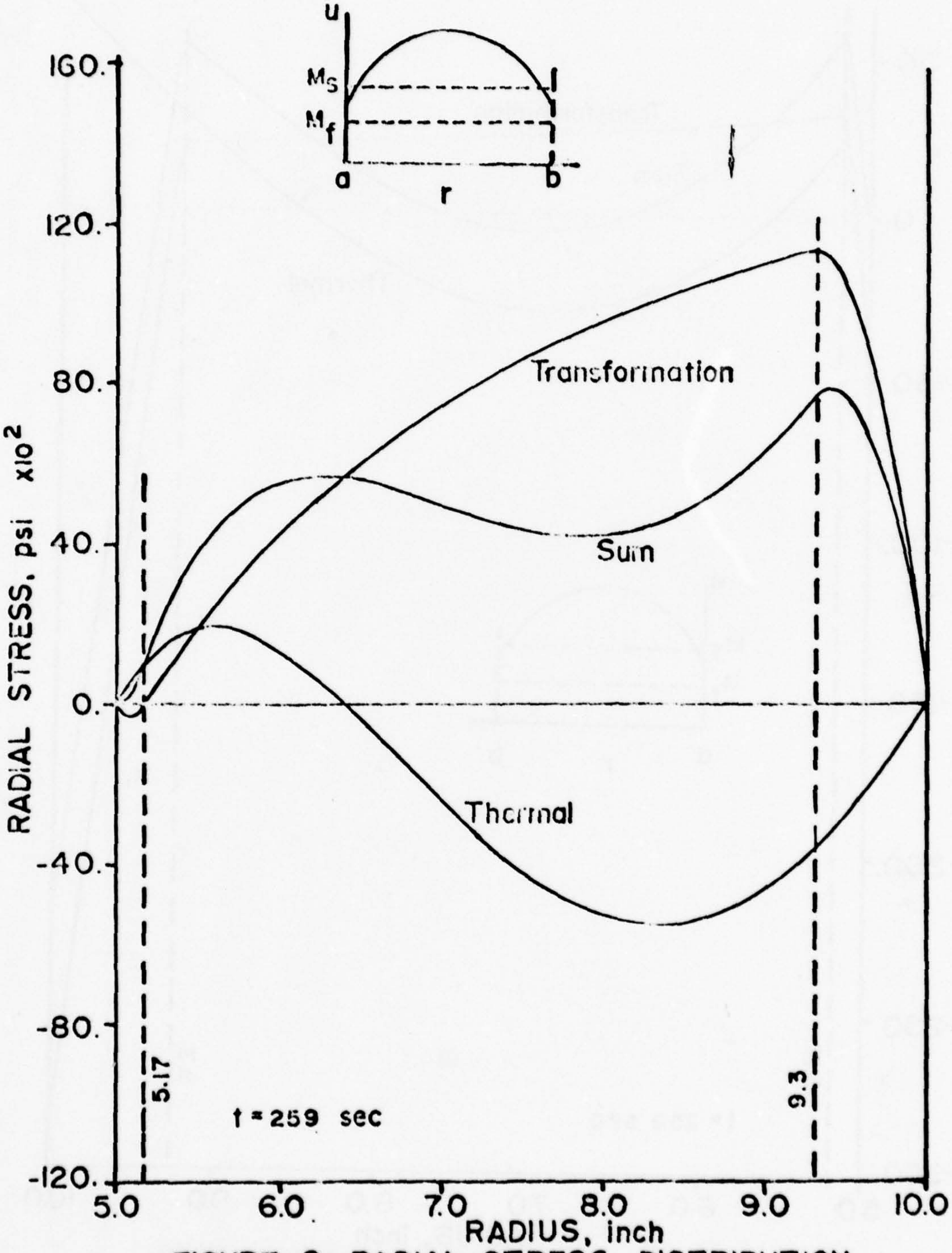


FIGURE 8. RADIAL STRESS DISTRIBUTION

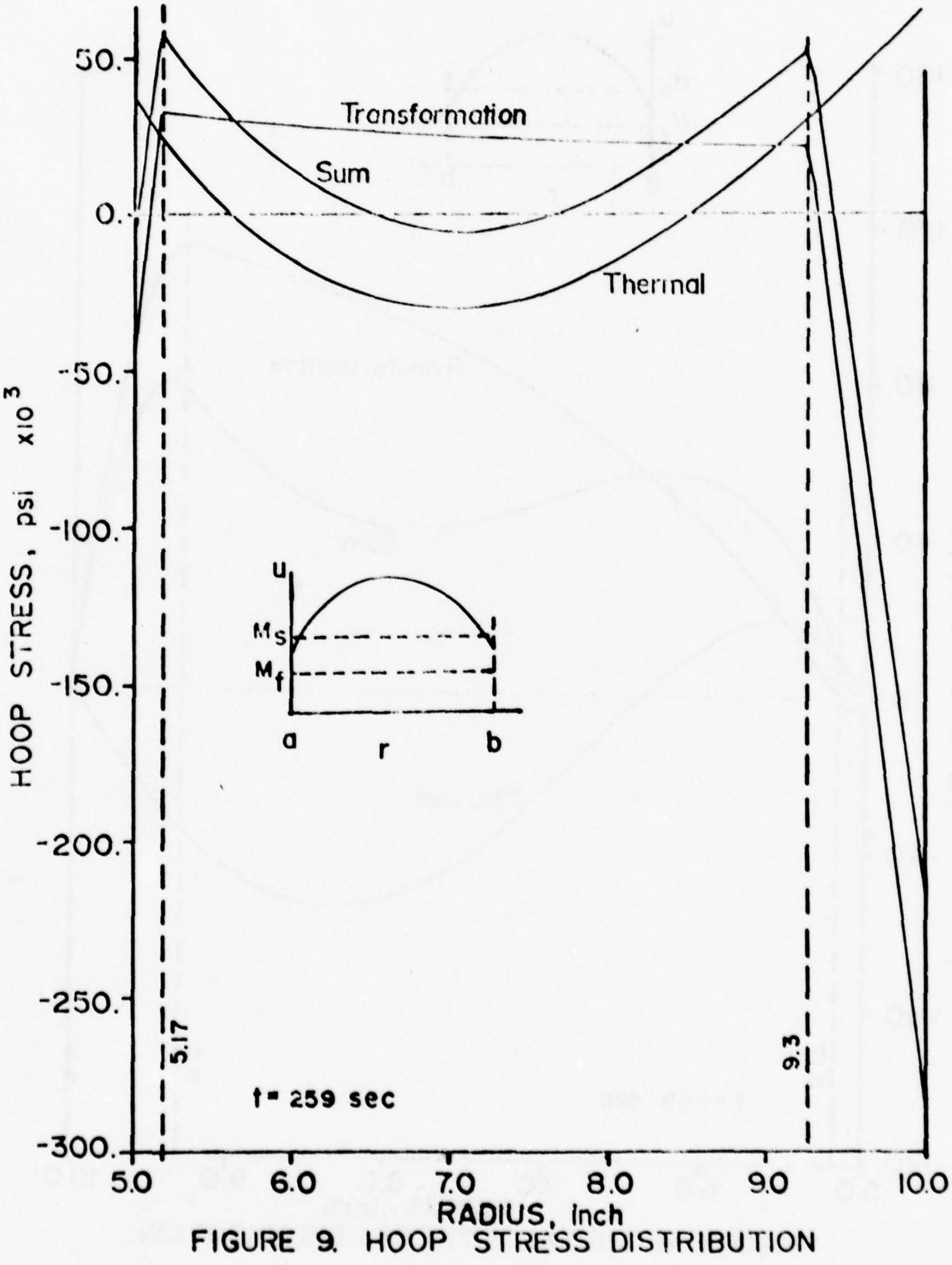


FIGURE 9. HOOP STRESS DISTRIBUTION

HYPERPARITY IN NONLINEAR SYSTEMS

Leon Kotin

Systems Analysis Division
Plans, Programs & Analysis Directorate
US Army Communications Research & Development Command
Fort Monmouth, NJ 07703

I. INTRODUCTION. It is surprising that even after the centuries during which ordinary differential equations have been studied, elementary techniques can still produce new and interesting results. In this paper, we extend the concept of parity to hyperparity by considering the effects of multiplying variables not only by -1, but by any root ω of unity. The principal consequence of these considerations will be to show the existence of periodic solutions of periodic differential systems, both linear and nonlinear.

II. SYMMETRY AND PARITY IN LINEAR DIFFERENTIAL EQUATIONS. We first consider the linear system

$$(1) \qquad \dot{X}(t) \equiv dX/dt = A(t)X, \qquad X(0) = I$$

where A and X are continuous rxr matrices and I is the identity matrix. In the following, ω denotes a root of unity, so that $\omega^n = 1$ for some integer n, and A^T denotes the transpose of A, for instance.

Theorem 1. If $\omega A^T(\omega t) \equiv -A(t)$, then $X^{-1}(t) = X^T(\omega t)$ for the principal fundamental solution X(t) of (1).

Proof. It is clear from (1) that $[X^T(\omega t)]^{\dot{}} = \omega X^T(\omega t)A^T(\omega t)$. Then $[X^T(\omega t)X(t)]^{\dot{}} = X^T(\omega t)[A(t) + \omega A^T(\omega t)]X(t)$. The hypothesis implies that this derivative is zero, so $X^T(\omega t)X(t) = \text{const.}$, and the initial value implies that this constant is I.

In the special case that $\omega = 1$, we conclude that if $A = -A^T$, then $X^{-1} = X^T$: i. e., if A is skew symmetric, then X is orthogonal. This result is well known. What is probably not known, however, is the

Corollary 2. If X(t) satisfies (1) with $A^T(-t) = A(t)$, then $X^{-1}(t) = X^T(-t)$.

Proof. With $\omega = -1$ in Theorem 1, the conclusion is immediate. It is easy to prove that the hypothesis on A is equivalent to: A is the sum of an odd skew symmetric matrix and an even symmetric matrix.

It follows further that since $X^{-1}(t) = X^{T}(\omega t)$, then $X(t) = X^{T-1}(\omega t) = X^{-1T}(\omega t) = X(\omega^{2}t)$; i. e., $X(t) = X(\omega^{2}t)$.

III. NONLINEAR SYSTEMS. We now apply similar considerations involving hyperparity (with $\omega^{n} = 1$) to nonlinear systems of ordinary differential equations.

In the sequel we shall assume that an existence and uniqueness theorem applies.

Theorem 3. Consider the system of differential equations

$$(2)\quad \dot{\underline{x}}_k(t) = \underline{f}_k(t,\underline{x}_1,\underline{x}_2,\dots,\underline{x}_n)$$
$$= \omega^{k+1}\underline{f}_k(\omega\, t,\ \omega^{n-1}\underline{x}_1, \omega^{n-2}\underline{x}_2,\dots,\ \omega\underline{x}_{n-1},\underline{x}_n)$$
$$(k = 1,\ 2,\ \dots,\ n)$$

with

$$(3)\qquad \underline{x}_k(0) = \underline{0} \qquad (k = 1,\ 2,\ \dots,\ n-1),$$

where $\dim \underline{x}_k \geq 0$. Then $\underline{x}_k(t) = \omega^{k}\underline{x}_k(\omega\, t)$, $k = 1,2,\dots,n$.

Proof. Define $\underline{\xi}_k(t) \equiv \omega^{k}\underline{x}_k(\omega\, t)$, $k = 1,2,\dots,n$. Then from condition (2) satisfied by each $\underline{f}_k$, it is not difficult to show that each $\underline{\xi}_k$ satisfies (2) and $\underline{\xi}_k(0) = \underline{x}_k(0)$. From the uniqueness of the solution, it follows that $\underline{\xi}_k(t) \equiv \underline{x}_k(t)$, which completes the proof.

By applying Theorem 3 to the case where $\dim \underline{x}_k = 0$ for all $k \neq n$ and then to the case where only $\underline{x}_{n-1}$ appears, we obtain the following corollaries.

Corollary 4. Let $\dot{\underline{x}} = \underline{f}(t,\underline{x}) \equiv \omega\underline{f}(\omega t,\underline{x})$. Then $\underline{x}(t) = \underline{x}(\omega t)$.

Corollary 5. Let $\dot{\underline{x}}(t) = \underline{f}(t,\underline{x}) = \underline{f}(\omega t, \omega\underline{x})$, $\underline{x}(0) = \underline{0}$. Then $\underline{x}(\omega t) = \omega\underline{x}(t)$.

We shall see that the case n = 2 is particularly important when periodicity is considered.

Corollary 6. Consider the system of two vector equations

$$\dot{\underline{x}} = \underline{f}(t,\underline{x},\underline{y}) \equiv \underline{f}(-t,-\underline{x},\underline{y})$$
$$\dot{\underline{y}} = \underline{g}(t,\underline{x},\underline{y}) \equiv -\underline{g}(-t,-\underline{x},\underline{y})$$

where $\underline{x}(t)$ satisfies the initial condition $\underline{x}(0) = \underline{0}$. Then $\underline{x}(t) = \underline{x}(-t)$, $\underline{y}(t) = \underline{y}(-t)$.

IV. PERIODIC SYSTEMS. **If, in addition to the conditions of parity in Corollary 6, $\underline{f}$ and $\underline{g}$ are periodic in t, we obtain a sufficient condition for the existence of a periodic solution. The remaining results and Corollary 6 have appeared in [3]. We shall assume that all solutions are extendable over the stated intervals.**

Theorem 7. Consider the system of two vector differential equations

$$(4) \qquad \dot{\underline{x}} = \underline{f}(t,\underline{x},\underline{y}) \equiv \underline{f}(-t,-\underline{x},\underline{y}) \equiv \underline{f}(t+\omega,\underline{x},\underline{y})$$

$$\dot{\underline{y}} = \underline{g}(t,\underline{x},\underline{y}) \equiv -\underline{g}(-t,-\underline{x},\underline{y}) \equiv \underline{g}(t+\omega,\underline{x},\underline{y})$$

with

$$\underline{x}(0) = \underline{x}(m\tau/2) = \underline{0}$$

for some integer m. Then

$$\underline{x}(t+m\tau) = \underline{x}(t), \quad \underline{y}(t+m\tau) = \underline{y}(t).$$

Proof. Define $\underline{\xi}(t) \equiv \underline{x}(t+m\tau)$, $\underline{\eta}(t) \equiv \underline{y}(t+m\tau)$; because of the periodicity of $\underline{f}$ and $\underline{g}$, $\underline{\xi}(t)$ and $\underline{\eta}(t)$ clearly satisfy (4). Now let $t = -m\tau/2$. Then $\underline{\xi}(-m\tau/2) = \underline{x}(m\tau/2) = -\underline{x}(-m\tau/2) = \underline{0}$ and $\underline{\eta}(-m\tau/2) = \underline{y}(m\tau/2) = \underline{y}(-m\tau/2)$, by Corollary 6. From the uniqueness of the solution, it follows that $\underline{\xi}(t) \equiv \underline{x}(t)$ and $\underline{\eta}(t) \equiv \underline{y}(t)$, which completes the proof.

For an autonomous system, the condition of periodicity in t is automatically satisfied.

Corollary 8. Consider the system

$$\dot{\underline{x}} = \underline{f}(\underline{x},\underline{y}) = \underline{f}(-\underline{x},\underline{y})$$

$$\dot{\underline{y}} = \underline{g}(\underline{x},\underline{y}) = -\underline{g}(-\underline{x},\underline{y})$$

where $\underline{x}(t)$ vanishes at two points: $\underline{x}(a) = \underline{x}(b) = \underline{0}$. Then $\underline{x}(t+2(b-a)) = \underline{x}(t)$, $\underline{y}(t+2(b-a)) = \underline{y}(t)$.

Proof. From the autonomy of the system, $\underline{x}(t+a)$ and $\underline{y}(t+a)$ are also solutions. *Now apply Theorem 7.*

The next special case of Theorem 7 is a generalization of a theorem of Demidovič [1], who showed that in the linear system

$$(5) \qquad \dot{\underline{x}} = A(t)\underline{x},$$

if A(t) is an odd function, skew symmetric and periodic, then all solutions are periodic. The condition of skew symmetry was eliminated by Epstein [2].

We now permit nonlinearity and conclude that if the system is merely periodic and odd in t, then all solutions are periodic. This follows from Theorem 7 by letting dim $\underline{x} = 0$ and choosing $m = 1$. Although it is an immediate corollary of Theorem 7, it is important enough to be favored with the title

Theorem 9. If

$$\dot{\underline{y}} = \underline{g}(t,\underline{y}) = -\underline{g}(-t,\underline{y}) = \underline{g}(t+\tau,\underline{y}),$$

then

$$\underline{y}(t+\tau) = \underline{y}(t).$$

V. THE SECOND-ORDER SCALAR DIFFERENTIAL EQUATION. We now further specialize Theorem 7 by considering second-order scalar equations. Some of the results in this section or their specializations appear scattered through the literature (cf.[4], p. 293;[5] , p.220;[6] ,p. 404).

Theorem 10. Consider the scalar equation

$$\ddot{x} = g(t,x,\dot{x}) = -g(-t,-x,\dot{x}) = g(t+\tau,x,\dot{x})$$

with $x(0) = x(m\tau/2) = 0$ for some integer m. Then $x(t+m\tau) = x(t)$.

Proof. Let $\dot{x} = y$ and apply Theorem 7 to the resulting system $\dot{x} = y$, $\dot{y} = g(t,x,y)$.

For an autonomous equation, the existence of any two zeros of a solution implies periodicity of that solution.

Corollary 11. If $\ddot{x} = g(x,\dot{x}) = -g(-x,\dot{x})$ and $x(a) = x(b) = 0$, then $x(t+2(b-a)) = x(t)$.

By reversing the roles of x and y we obtain analogous results involving zeros of the derivative.

Theorem 12. If $\ddot{y} = f(t,\dot{y},y) = f(-t,-\dot{y},y) = f(t+\tau,\dot{y},y)$ and $\dot{y}(0) = \dot{y}(m\tau/2) = 0$ for some integer m, then $y(t+m\tau) = y(t)$.

Proof. Let $\dot{y} = x$ and apply Theorem 7 to the resulting system $\dot{x} = f(t,x,y)$, $\dot{y} = x$.

Corollary 13. If$\ddot{y} = f(\dot{y},y) = f(-\dot{y},y)$ and $\dot{y}(a) = \dot{y}(b) = 0$, then $y(t+2(b-a)) = y(t)$.

REFERENCES

[1] Demidovič, B.P., On some properties of some characteristic exponents of a system of ordinary linear differential equations with periodic coefficients. Uch. Zap., Mosk. Gos. Univ. No. 163, Mat. 6, 123-136 (1952).

[2] Epstein, I.J., Periodic solutions of systems of differential equations. Proc. Am. Math. Soc. 13, 690-694 (1962).

[3] Kotin, L., Solutions of Systems of periodic differential equations. J. Math. Anal. Appl. 8, 52-56 (1964).

[4] Krasnoselskii, M.A., "Positive Solutions of Operator Equations." Noordhoff, Groningen, 1964.

[5] Pliss, V.A., "Nonlocal Problems in the Theory of Oscillations," Academic Press, New York, 1966.

[6] Sansone, G. and Conti, R., "Non-linear Differential Equations," Rev. Ed., MacMillan, New York, 1964.

ASYMPTOTIC SOLUTIONS TO A STABILITY PROBLEM

David A. Peters
Department of Mechanical Engineering
Washington University
Saint Louis, Missouri 63130 USA

Julian J. Wu
U.S. Army Armament Research and Development Command
Benet Weapons Laboratory, LCWSL
Watervliet Arsenal
Watervliet, New York 12189 USA

ABSTRACT. This paper is concerned with the lateral stability of a free-flying column subjected to an axial thrust with directional control. The stability curve (i.e., eigenvalue vs. thrust, in the neighborhood of zero eigenvalues) and the associated eigenfunctions of this problem have not been fully understood. This paper uses asymptotic expansions to examine closely, for all values of the thrust directional-control parameter, both the intersection of the eigenvalue curves with the zero branch and the associated eigenfunctions of zero and nearly zero eigenvalues. Several analytical proofs are provided substantiating previous numerical findings.

1. INTRODUCTION.

A slender, uniform column of Euler-Bernoulli type, subjected to a thrust at one end and travelling freely in the axial direction, is the simplest structural model conceivable for a *flexible rocket*, missile, or space-craft. The lateral disturbance of such a column in non-dimensionalized form is governed by the following differential equation (see, for example, reference [1])

$$u'''' + Q(xu')' + \ddot{u} = 0 \qquad \text{(1a)}$$

with boundary conditions

$$u''(0) = 0, \quad u''(1) = 0, \quad u'''(o) = 0 \qquad \text{(1b,1c,1d)}$$

and

$$u'''(1) - K_\theta Q u'(1) = 0 \qquad \text{(1e)}$$

where $u = u(x,t)$ denotes the lateral disturbance of the column from its equilibrium position as a function of the spatial coordinates x and the time t; Q is the thrust at the end; and K_θ is a feedback control parameter indicating the angle between the direction of Q and the tangent of the column at the end. A prime (') denotes differentiation with respect to x, and a dot ($\cdot$), differentiation with respect to t. The boundary conditions simply state the bending moments and shear forces that must be satisfied at the ends. It is known in practical design that a suitable choice of K_θ in Eq. (1e) will improve the stability performance of the structure.

For vibrations and quasi-dynamic stability problems, one can eliminate the time variable by assuming that

$$u(x,t) = u(x)e^{\lambda t} \tag{2}$$

Eq. (1a) then becomes

$$u'''' + Q(xu')' + \lambda^2 u = 0 \tag{1a'}$$

and the initial conditions do not enter into the problem. "Eqs. (1')" will be used to refer to Eqs. (1) with Eq. (1a) replaced by Eq. (1a'). For most of the results of this paper Eqs. (1') will be used. However, in the case of repeated eigenvalues with identical eigenfunctions, Eqs. (1) must be used to find another independent solution. This will be discussed in Section 5.

It is clear from Eq. (2) that the parameter λ dictates the stability nature of the problem. A purely imaginary λ indicates stable vibrations; a purely real and positive λ, instability of divergence and a complex λ with positive real part, instability of flutter. Since λ appears only as λ^2 in Eqs. (1), it is equally true that a real negative λ^2 indicates stable vibrations; a real positive λ^2, instability of divergence; and a complex λ^2, instability of flutter.

Eqs. (1') defines a problem with two somewhat unusual features, both pertaining to the boundary conditions, that have caused considerable difficulty in seeking a full understanding of the solutions. First, the problem is nonself-adjoint except for the special case where $K_\theta = -1$. Second, the solution always involves zero eigenvalues. Many puzzling aspects associated with these features will be considered in this paper.

In order to place this investigation in proper perspective, a brief account is given of previous work which has lead to the present results.

The vibrations of a free-free column without axial force can be found in many text books (see, for example, reference [2]). Usually only one zero eigenvalue (in terms of λ^2 in Eqs. (1')) is recorded, although it is obvious that there are two zero eigenvalue solutions, one corresponding to rigid body translation and one to rigid body rotation. This fact turns out to be significant in seeking a full understanding of the stability problem when an axial thrust is present. Silverberg [3] appears to be first in attempting to solve the problem of a free flying column subjected to an axial thrust fixed in the direction of the undisturbed axis - a special case of Eqs. (1) in which $K_\theta = -1$. Silverburg did not realize, however that one of the two zero eigenvalues becomes a divergence (buckling) branch as soon as the thrust Q becomes non-zero; and he obtained a stability criterion from the first stable (vibrations) branch of eigenvalues.

This branch begins at $Q = 0$ and varies with Q until a "buckling" thrust is reached for which the eigenvalue of the first stable branch becomes zero. Silverberg also obtained analytically this "buckling" load as a zero of a Bessel function. This particular value of thrust is now known to be significant only for $K_\theta > 0$, but it has no meaning as a stability criterion for $K_\theta < 0$. Beal [4] used Galerkin's method and performed a thorough analysis for the general problem including pulsating thrusts. For the case of a constant thrust and $K_\theta = 0$, he obtained the coalescence branches and the first critical thrust of flutter. The Galerkin technique gave two zero eigenvalues for all values of Q at $K_\theta = 0$. Beal concluded from physical reasoning that one eigenvalue was associated with a rigid body translation mode; and the other with a rigid body rotation accompanied by translation. The fact that the rigid body rotation [$u'(x) = \text{constant} \neq 0$] can not satisfy the differential equation (1a') was not addressed. Beal also recognized that for $K_\theta \neq 0$ there are repeated zero eigenvalues at certain values of thrust Q. Beal hypothesized from physical reasoning that for $K_\theta \neq -1$ the two mode shapes at the intersection were: 1) a rigid body translation, and 2) a combination of rigid body rotation and the Silverberg buckling shape. Again however, he did not comment on the failure of the second "mode" to satisfy the eigenvalue equation. At about the same time in Russia, Feodos'ev [5], used a method of truncated polynomials to obtain the same coalescence branches as reported by Beal. Similar results for $K_\theta = 0$ were reported by Matsumoto and Mote [6] who used finite elements in conjunction with an extended Hamilton's principle. Wu [7,8] used finite elements adjoint variational formulations and presented data which showed that although $K_\theta > 0$ has a stabilizing effect, $K_\theta < 0$ is actually destabilizing in the interval $0<Q<2.60\pi^2$. Wu's earlier attempt to resolve the dilemma on the rigid body rotation failed due to a numerical error as was pointed out by Sundararamiah and Johns [9]. The numerical data for a wide range of K_θ values were subsequently corrected [10]. For the special case $K_\theta = 0$, however, the conclusions drawn [10] from the numerical analysis were still unsatisfactory with respect to the repeated eigenvalues and associated eigenfunctions. Numerical results reported to date [4,10,11] also indicate that the eigenvalue curves all approach zero at identical values of Q (i.e., independent of K_θ). The analytical proof of this offered in [10] was also not satisfactory due to the fact that one of the boundary conditions was not satisfied rigorously except for the case where $K_\theta = -1$.

In this paper, the mode shape solutions at zero frequency are first summarized for $Q = 0$ and for $K_\theta = -1$ in Section 2. In Section 3, the more elusive cases of $K_\theta \neq -1$ ($K_\theta = 0$ and $K_\theta \neq 0$) are treated by the method of asymptotic expansion. Analytical proofs are provided supporting the earlier numerical findings. These are accomplished with the help of other asymptotic expansions and are presented in Section 4. The meaning of the repeated eigenvalues and their associated eigenfunctions is discussed in Section 5. Numerical calculations that verify the convergence of the expansion formulas are provided in Section 6 and the conclusions of this investigation are summarized in Section 7.

2. SOLUTIONS AT ZERO EIGENVALUES. It will be worthwhile first to see what can be concluded at $\lambda = 0$ without the use of asymptotic expansions. With $\lambda = 0$, Eqs. (1') become

$$u'''' + Q(xu')' = 0 \tag{3a}$$

$$u''(0) = 0, \quad u''(1) = 0, \quad u'''(0) = 0 \tag{3b,3c,3d}$$

$$u'''(1) - K_\theta Qu'(1) = 0 \tag{3e}$$

Some simple observations can be made from Eqs. (3).

1. For any value of Q and of K_θ, $u(x) = u(0)$ is a solution where $u(0)$ is an arbitrary constant.

2. For $Q = 0$ and for any value of K_θ, another solution is $u(x) = v(0)x$ where $v(0)$ is an arbitrary constant quite independent of $u(0)$.

3. For $Q \neq 0$ and for any value of K_θ, $u(x) = v(0)x$, with $v(0) \neq 0$, can not be a solution to Eqs. (3).

Since at $Q = 0$, there is a double zero eigenvalue and two independent eigenfunctions (mode shapes), one could expect that two branches of the eigenvalue curve would emanate from $Q = 0$ as Q increases. One of these is the identically zero branch that has the mode $u(x) = u(0)$. We wish to determine the behavior of the second zero branch as well as the behavior of any other branches that might cross the zero branch at other values of Q for various K_θ. A plot of λ (real and imaginary) vs. Q from numerical calculations [10] is shown in Figure 1. It clearly shows that $\lambda = 0$ intersections do exist for all non zero values of K_θ.

The following is also easily observed from Eqs. (1). Since λ only appears as λ^2 in (1), the functional relationship between λ^2 and Q can be written as $\lambda^2 = f(Q)$. Thus $d\lambda/dQ$ goes to infinity as λ goes to zero. In other words, if the λ vs. Q curve is to approach $\lambda = 0$, it must do so perpendicularly.

Now we proceed to solve Eqs. (3). Eq. (3a) can be integrated once

$$u''' + Qxu' = c = 0 \tag{4}$$

The constant of integration is $c = 0$ because of Eq. (3d). Due to Eqs. (3e) and (4), one has

$$Q(1+K_\theta)u'(1) = 0 \tag{5}$$

Now, let

$$u'(x) = v(x) \tag{6a}$$

$$u(x) = \int_0^x v(\xi)d\xi + u(0) \tag{6b}$$

Eqs. (3) can be written in terms of v(x) as

$$v'' + Qxv = 0 \tag{7a}$$

$$v'(0) = 0, \quad v'(1) = 0 \tag{7b,7c}$$

$$Q(1+K_\theta)v(1) = 0 \tag{7d}$$

Since the solutions for $Q = 0$ are already known, only $Q \neq 0$ will be considered here. Two cases of Eq. (7d) need to be considered: $K_\theta = -1$ and $K_\theta \neq -1$.

For $K_\theta = -1$, Eq. (7d) is satisfied without any restriction on v(1). Thus the non-trivial solution of Eqs. (7) will yield $\lambda = 0$ eigenfunctions of Eq. (1') that are independent of $u(x) = u(0)$. These solutions of Eqs. (7a) are known to be Airy functions [12] and can be written as

$$v(x) = (Q^{1/3} x)^{1/2} [A J_{1/3}(\tfrac{2}{3} Q^{1/2} x^{3/2}) + B J_{-1/3} (\tfrac{2}{3} Q^{1/2} x^{3/2})] \tag{8}$$

where J denotes the Bessel function of the first kind and A, B are constants to be determined by boundary conditions. One concludes from the boundary conditions (7b) and (7c) that $A = 0$ and that Q must assume one of the discrete values denoted by $\hat{Q}_j$, $j = 0, 1, 2, \ldots$ such that

$$J_{2/3} (\tfrac{2}{3} \hat{Q}_j^{1/2}) = 0 \tag{9a}$$

the first five of such $\hat{Q}_j$'s are (Figure 1, also see [10])*

$$\frac{\hat{Q}_j}{\pi^2} = 0; \quad 2.598; \quad 9.722; \quad 21.35; \quad 37.49 \tag{9b}$$

*Throughout this paper, analytical proofs substantiating previous numerical findings will be noted by a reference to the appropriate figure.

Thus the nontrivial solution of Eqs. (7) at $Q = \hat{Q}_j$ can be written as

$$v(x) = B\,(Q^{1/3}\,x)^{1/2}\,J_{-1/3}\left(\frac{2}{3}Q^{1/2}x^{3/2}\right)$$

$$v(x) = v(1)\phi(x) \tag{10a}$$

$$\phi(x) = \Phi(Q^{1/3}\,x)/\Phi(Q^{1/3}) \tag{10b}$$

$$\Phi(y) = y^{1/2}\,J_{-1/3}\left(\frac{2}{3}y^{3/2}\right) \tag{10c}$$

and v(1) is an arbitrary constant. Now the solution of Eqs. (1') for $K_\theta = -1$ and at $Q = \hat{Q}_j$, from Eq. (6b), is then

$$u(x) = u(0) + v(1)\int_0^x \phi(\xi)d\xi \tag{11}$$

where $\phi(x)$, as defined in Eq. (10b), is a function of Q as well as of x. It is also observed that $\phi(1) = 1$.

Since v(1) is quite independent of u(0), both being arbitrary constants, u(x) of Eq. (11) is an eigenfunction independent of u(x) = u(0). It is then clear that for $K_\theta = -1$ and at $Q = \hat{Q}_j$ of Eqs. (9), there exists a double zero eigenvalue with two independent eigenfunctions*

$$u(x) = 1 \tag{12a}$$

$$u(x) = \int_0^x \phi(\xi)d\xi \tag{12b}$$

where $\phi(x)$ is defined in Eq. (10b).

For $Q = \hat{Q}_o = 0$, Eqs. (12) reduce to

$$u(x) = 1 \tag{13a}$$

$$u(x) = x \tag{13b}$$

which agrees with the special case for Q = 0 and for arbitrary K_θ observed earlier.

*From here on, the normalized form of an eigenfunction will be used.

Now for $K_\theta \neq -1$ in Eqs. (7), and with $Q \neq 0$, Eq. (7d) demands that $v(1) = 0$. Thus from the analysis above, Eqs. (7) only admits the trivial solution $v(x) = 0$. This suggests that for $K_\theta \neq -1$ even at those $\hat{Q}_j$ of Eqs. (9), the only solution is $u(x) = u(0)$. Thus, for $Q \neq 0$ and $K_\theta \neq -1$, the intersection of two eigenvalue branches at $\lambda = 0$ has only one associated eigenvector. How do the eigenvectors transition to a rigid body translation as λ approaches zero? An answer to this question is provided by the asymptotic expansions.

3. ASYMPTOTIC SOLUTIONS. We are interested in finding the manner in which the eigenvalue curves of Eqs. (1') intersect the zero branch (for various K_θ) and the nature of the associated eigenfunctions near those crossings. It is already known that $Q = 0$ is such a crossing for all values of K_θ and that two independent eigenfunctions exist there, Eqs. (13). For $K_\theta = -1$, the crossings are at those $\hat{Q}_j$'s of Eqs. (9) and two independent eigenfunctions exist there, Eqs. (12). Thus in this section, we consider the cases $Q \neq 0$ and $K_\theta \neq -1$.

Since an eigenfunction $u(x)$ is a function of its eigenvalue λ^2, for $|\lambda^2|<<1$, one can expand $u(x)$ and Q as power series in λ^2:

$$u(x) = u_0(x) + \lambda^2 u_1(x) + \lambda^4 u_2(x) + \ldots \qquad (14a)$$

$$Q = Q_0 + \lambda^2 Q_1 + \lambda^4 Q_2 + \ldots \qquad (14b)$$

The parameter K_θ is considered fixed for the moment. One can obtain equations for the n-th order in λ^2, $n = 0, 1, 2, \ldots$, by substitution of Eqs. (14) into Eq. (1')

$$u_n'''' + Q_0(xu_n')' = -u_{n-1} - \sum_{i=1}^{n} Q_i(xu'_{n-i})' \qquad (15a)$$

$$u_n''(0) = 0, \quad u_n''(1) = 0, \quad u_n'''(0) = 0 \qquad (15b,15c,15d)$$

$$u_n'''(1) - K_\theta Q_0 u_n'(1) = \sum_{i=1}^{n} K_\theta Q_i u'_{n-i} \qquad (15e)$$

where, for $n = 0$, the right-hand-side term of Eq. (15a) is zero. It is noted that Eqs. (15) is recurrent meaning that each set of equations in $u_n(x)$ depends on the solution of $u_{n-1}(x)$. Each recurrent Q_i is found by enforcing the boundary conditions on the solution u_{i+1}. Through a procedure similar to that used in the previous section, Eqs. (15) are transformed into an equivalent set

$$v_n'' + Q_0 x v_n = -\int_0^x u_{n-1}(\xi)d\xi - \sum_{i=1}^{n} Q_i x v_{n-i} \qquad (16a)$$

$$v_n'(0) = 0, \quad v_n'(1) = 0 \tag{16b,16c}$$

$$\sum_{i=0}^{n} Q_i(1+K_\theta)v_{n-i}(1) = -\int_0^1 u_{n-1}(\xi)d\xi \tag{16d}$$

$$u_n'(x) = v_n(x) \tag{17a}$$

$$u_n(x) = \int_0^x v_n(\xi)d\xi + u_n(0) \tag{17b}$$

Only the zeroth, first, and second order equations will be considered here. For n = 0, Eqs. (16) become

$$v_0'' + Q_0 x v_0 = 0 \tag{18a}$$

$$v_0'(0) = 0, \quad v_0'(1) = 0 \tag{18b,18c}$$

$$Q_0(1+K_\theta)v_0(1) = 0 \tag{18d}$$

The set of zeroth order equation (18) is identical to Eqs. (7) as expected. Since $Q \neq 0$, $K_\theta \neq -1$, the only solution to Eqs. (18) is

$$v_0(x) = 0 \tag{19a}$$

$$u_0(x) = 1 \tag{19b}$$

where the mode is normalized with u(0) = 1.

We now proceed to the next order equations. For n = 1, one has

$$v_1'' + Q_0 x v_1 = -\int_0^x u_0(\xi)d\xi = -x \tag{20a}$$

$$v_1'(0) = 0, \quad v_1'(1) = 0 \tag{20b,20c}$$

$$Q_0(1+K_\theta)v_1(1) = -1 \tag{20d}$$

To solve Eqs. (20), let

$$w_1(x) = v_1(x) + \frac{1}{Q_0} \tag{21a}$$

$$v_1(x) = w_1(x) - \frac{1}{Q_0} \tag{21b}$$

Substitution of Eq. (21b) in Eqs. (20), gives

$$w_1'' + Q_0 x w_1 = 0 \tag{22a}$$

$$w_1'(0) = 0, \quad w_1'(1) = 0 \tag{22b,22c}$$

$$w_1(1) = \frac{K_\theta}{Q_0(1+K_\theta)} \tag{22d}$$

Two cases will be considered separately: $K_\theta \neq 0$ and $K_\theta = 0$.

First, consider the case $K_\theta \neq 0$. From the non-trivial solution to Eqs. (7) in the previous section, one can immediately write down the non-trivial solution to Eqs. (20) as

$$\begin{aligned} w_1(x) &= w_1(1)\,\phi(x) \\ &= \frac{K_\theta}{Q_0(1+K_\theta)}\,\phi(x) \end{aligned}$$

where $\phi(x)$ is given in Eq. (10b) with Q replaced by Q_0. It is clear that Q_0 may take on any of the $\hat{Q}_j$'s of Eqs. (9) excluding $\hat{Q}_0 = 0$. Now,

$$\begin{aligned} u_1(x) &= \int_0^x v_1(\xi)d\xi + u_1(0) = \int_0^x w_1(\xi)d\xi - \frac{x}{Q_0} + u_1(0) \\ &= \frac{K_\theta}{Q_0(1+K_\theta)} \int_0^x \phi(\xi)d\xi - \frac{x}{Q_0} + u_1(0) \end{aligned}$$

The constant $u_1(0)$ must vanish to preserve the normalization $u(0) = 1$. The combined zeroth and first order solution is therefore

$$\begin{aligned} u(x) &= u_0(x) + \lambda^2 u_1(x) \\ &= 1 + \lambda^2 \left[\frac{K_\theta}{Q_0(1+K_\theta)} \int_0^x \phi(\xi)d\xi - \frac{x}{Q_0} \right] \end{aligned} \tag{23}$$

The solution given by Eq. (23) provides the following information for $K_\theta \neq -1$:

1. For $|\lambda^2|<<1$, non-trivial solutions exist at those same $\hat{Q}_j$ at which the second branch crosses the zero branch in the case $K_\theta = -1$.

2. As long as λ^2 is not identically zero, Eq. (23) provides an eigenfunction, independent of $u(x)$ = constant. This independent eigenfunction (mode shape) consists of a finite rigid body translation, a bending term of an integral of an Airy function proportional to λ^2, and a rigid body rotation, also proportional to λ^2. These three terms are inseparable as an independent eigenfunction since they all are proportional to the arbitrary constant $u(0)$ as shown in Eq. (23).

3. As λ^2 goes to zero, Eq. (23) reduces smoothly to a rigid body translation. This again indicates that the double zero eigenvalue that exists at those $\hat{Q}_j$'s of Eqs. (9) has only one eigenfunction. In order to generate another independent function, one must use the concept of a "Jordon vector". This will be discussed in Section 5.

We have obtained an expression for the non-trivial mode shape, Eq. (23), near $\lambda^2 = 0$, for $K_\theta \neq -1$, $K_\theta \neq 0$ and Q near $\hat{Q}_j$ of Eqs. (9). This expression, however, is incomplete without a known Q_1 in the relationship between λ^2 and Q:

$$Q = Q_0 + \lambda^2 Q_1 \tag{24}$$

It is also clear from this equation that the curvature of a Q vs. λ in the neighborhood of $\lambda = 0$ is given by

$$\frac{d^2Q}{d\lambda^2} = 2Q_1 \tag{25}$$

But, in order to obtain Q_1, one needs the second order equation From Eqs. (16),

$$v_2'' + Q_0 x v_2 = -\int_0^x u_1(\xi)d\xi - Q_1 x v_1 - Q_2 x v_0 \tag{26a}$$

$$v_2'(0) = 0, \quad v_2'(1) = 0 \tag{26b,26c}$$

$$\int_0^1 u_1(\xi)d\xi + Q_0(1+K_\theta)v_2(1) + Q_1(1+K_\theta)v_1(1)$$
$$+ Q_2(1+K_\theta)v_0(1) = 0 \tag{26d}$$

From previous results, Eqs. (19), (23) after normalization

$$u_0(x) = 1 \quad , \quad u_1(x) = \frac{1}{Q_0}\left[\frac{K\theta}{1+K\theta}\int_0^x \phi(\xi)d\xi - x\right] \qquad \text{(27a,27b)}$$

$$v_0(x) = 0 \quad , \quad v_1(x) = \frac{1}{Q_0}\left[\frac{K_\theta}{1+K_\theta}\,\phi(x) - 1\right] \qquad \text{(27c,27d)}$$

Now let

$$v_2 = w_2 - \frac{Q_1}{Q_0}\,v_1 \qquad (28)$$

one has, from Eqs. (26),

$$w_2'' + Q_0 x w_2 = -\int_0^x u_1(\xi)d\xi + \frac{Q_1}{Q_0}\,v_1''$$

$$w_2'(0) = 0, \quad w_2'(1) = 0$$

$$w_2(1) = v_2(1) + \frac{Q_1}{Q_0}\,v_1(1)$$

where, from Eqs. (20) and (27d)

$$v_1'' = -x-Q_0 x v_1 = -\frac{K_\theta}{1+K_\theta}\,x\phi(x)$$

$$v_2(1) = -\frac{1}{Q_0(1+K_\theta)}\int_0^1 u_1(\xi)d\xi - \frac{Q_1}{Q_0}\,v_1(1)$$

Therefore,

$$w_2'' + Q_0 x w_2 = -\int_0^x u\ (\xi)d\xi - \frac{Q_1}{Q_0}\left(\frac{K_\theta}{1+K_\theta}\right)x\ \phi(x) \qquad \text{(29a)}$$

$$w_2'(0) = 0, \quad w_2'(1) = 0 \qquad \text{(29b,29c)}$$

$$w_2(1) = -\frac{1}{Q_0(1+K_\theta)}\int_0^1 u_1(\xi)d\xi \qquad \text{(29d)}$$

Let

$$w_2(x) = \phi(x)\psi(x) \tag{30a}$$

$$w_2'(x) = \phi'\psi + \phi\psi' \tag{30b}$$

$$w_2''(x) = \phi''\psi + 2\phi'\psi' + \phi\psi'' \tag{30c}$$

From Eqs. (29), one has

$$\psi(\phi'' + Q_0 x\phi) + 2\phi'\psi' + \psi''\phi = - \int_0^x u_1(\xi)d\xi - \frac{Q_1}{Q_0}\left(\frac{K_\theta}{1+K_\theta}\right)x\phi(x)$$

or

$$2\phi'\psi' + \psi''\phi = - \int_0^x u\ (\xi)d\xi - \frac{Q_1}{Q_0}\left(\frac{K_\theta}{1+K_\theta}\right)x\phi(x) \tag{31}$$

since

$$\phi'' + Q_0 x\phi = 0$$

by definition of the function $\phi(x)$ in Eq. (10b). From Eq. (31),

$$\frac{d}{dx}(\phi^2\psi') = - \phi(x)\int_0^x u_1(\xi)d\xi - \frac{Q_1}{Q_0}\left(\frac{K_\theta}{1+K_\theta}\right)x\phi^2(x)$$

or

$$\phi^2\psi' = - \int_0^x \phi(\xi)\left(\int_0^\xi u_1(\eta)d\eta\right)d\xi - \frac{Q_1}{Q_0}\left(\frac{K_\theta}{1+K_\theta}\right)\int_0^x \xi\phi^2(\xi)d\xi + c$$

Since $\psi'(0) = 0$ and $\psi'(1) = 0$ are results of $w_2'(0) = w_2'(1) = 0$ and $\phi'(0) = \phi'(1) = 0$, one concludes that $c = 0$ and that

$$Q_1 = - \frac{Q_0(1+K_\theta)}{K_\theta} \quad \frac{\int_0^1 \phi(\xi)\left(\int_0^\xi u_1(\eta)d\eta\right)d\xi}{\int_0^1 \xi\phi^2(\xi)d\xi}$$

When $u_1(\eta)$ in the equation above is replaced by Eq. (27b), one has

$$Q_1 = \frac{1}{\delta}\left[\frac{1+K_\theta}{K_\theta}\int_0^1 \frac{x^2}{2}\phi(x)dx - \int_0^1 \phi(x)\int_0^x \int_0^\xi \phi(\eta)d\eta d\xi dx\right] \tag{32a}$$

where

$$\delta = \int_0^1 x\phi^2(x)dx \tag{32b}$$

Eqs. (23), (24) and (32) form the complete set of asymptotic solutions of eigenvalues and eigenfunctions for Q near $\hat{Q}_j$. These solutions will be compared with finite element solutions in Section 6.

We shall now go back to Eqs. (20) for the case $K_\theta = 0$. The solution yields $w_1(x) = 0$; $v_1(x) = -1/Q_0$ for all non-zero values at Q_0, not just the $\hat{Q}_j$. The combined zeroth and first order solution to Eqs. (1) is thus

$$u(x) = u_0(x) + \lambda^2 u_1(x)$$

or

$$u(x) = 1 - \frac{\lambda^2}{Q} x \tag{33}$$

Eq. (33) suggests that for $K_\theta = 0$, there exists eigenfunctions independent of $u(x) = 1$ for some arbitrarily small λ^2. However, we shall prove that this arbitrarily small λ^2 has to be zero for all values of Q.

Since K_θ is expected to be a continuous function of λ^2, its first expansion term about $K_\theta = 0$ in terms of λ^2 must be in the form of

$$K_\theta = \lambda^2 K_1 \tag{34}$$

for Q held constant. In a latter section, it will be shown that K_1 is never zero and thus K_θ approaching zero requires that λ^2 is also approaching zero. Thus Eq. (33) reduces to

$$u(x) = 1 \tag{33'}$$

Hence, for $K_\theta = 0$, and for all nonzero values of Q one has a situation of double zero eigenvalues with only one eigenfunction. The second zero eigenvalue branch does not cross $\lambda = 0$, however, it coincides with it.

4. OTHER RESULTS BY ASYMPTOTIC EXPANSIONS. It will be proved in this section that for $|\lambda^2| \ll 1$ and $\hat{Q}_j < \hat{Q} < \hat{Q}_{j+1}$, where $\hat{Q}_j$ are given by Eqs. (9), λ^2 is negative (stable vibrations) for a positive K_θ and λ^2 is positive (divergence instability) for a negative K_θ if j = 0, 2, 4, ... The sign of λ^2 will be reversed if i = 1, 3, 5, ... (Figure 1). It will be proved also that K_1 in Eq. (34) can not be zero for all nonzero values of Q. Two asymtotic expansions will be used here.

First, let $Q_0 = 0$ which leaves

$$Q = \lambda^2 Q_1 + \lambda^4 Q_2 + \ldots \tag{35}$$

The zeroth order equations (16) are

$$v_0'' = 0 \tag{36a}$$

$$v_0'(0) = 0, \quad v_0'(1) = 0 \tag{36b,36c}$$

The solution to Eqs. (36) is

$$v_0(x) = u_0'(x) = v_0(0), \text{ a constant} \tag{37a}$$

$$u_0(x) = v_0(0)x + 1 \tag{37b}$$

The mode shape $u_0(x)$ has been normalized so that $u(0) = u_0(0) = 1$. The constant $v_0(0)$ is yet to be determined. The first order equations, from Eqs. (16), are

$$v_1'' = - \int_0^x u_0(\xi)d\xi - Q_1 x \tag{38a}$$

$$v_1'(0) = 0, \quad v_1'(1) = 0 \tag{38b,38c}$$

$$Q_1(1+K_\theta)v_0(1) = - \int_0^1 u_0(\xi)d\xi \tag{38d}$$

From Eq. (37b), one obtains

$$v_1'' = - \frac{1}{2} v_0(0)x^2 - [1 + Q_1 v_0(0)]x \tag{39a}$$

$$v_1'(0) = 0, \quad v_1'(1) = 0 \tag{39b,39c}$$

$$Q_1(1+K_\theta)v_0(0) + \frac{1}{2} v_0(0) = -1 \tag{39d}$$

Eq. (39a) can be integrated directly. From the boundary conditions (39b)-(39d) one has, for non-trivial solution of $v_1(x)$,

$$Q_1 = - \frac{1}{6K_\theta} \tag{40a}$$

$$Q = \lambda^2 Q_1 = - \frac{\lambda^2}{6K_\theta} \tag{40b}$$

and

$$v_0(0) = \frac{6K_\theta}{1-2K_\theta} \tag{40c}$$

Hence the mode shape in this case (Q near zero) is uniquely defined, from Eq. (37b), as

$$u_0(x) = (\frac{6K_\theta}{1-2K_\theta})x + 1 \tag{41}$$

Equation (41) shows that the mode approaches a combination of translation and rotation as Q approaches zero. For $K_\theta = \frac{1}{2}$, it approaches a pure rotation; and for $K_\theta = 0$, it approaches a pure translation. For the latter case, equation (40b) shows that λ equals zero for all Q which implies that the branch approaches horizontally (i.e. coincides with the $\lambda = 0$ axis). Therefore, the rigid body rotation mode at Q = 0 is an isolated point for $K_\theta = 0$; and the mode changes to rigid body translation for arbitrarily small Q. Since Q is always real and is assumed here to be positive, one concludes from Eqs. (40) that for Q near zero, ε positive K_θ produces a negative λ^2, and a negative K_θ produces a positive λ^2 (Figures 1 and 2).

Next, we shall fix Q and expand K_θ in λ^2 about $K_\theta = 0$ in order to determine the behavior of the branches and mode shapes as K_θ approaches zero. Thus,

$$K_\theta = \lambda^2 K_1 + \lambda^4 K_2 + \ldots. \tag{42}$$

Substitution of Eqs. (14) and (42) into (1') yields equations of various orders in λ^2.

Since the zeroth and the first order equations are identical with Eq. (18) and (20) respectively (with zero substituting for K_θ and Q substituting Q_0 in these equations), one can simply write down the solutions from Eqs. (19) and (33):

$$v_0(x) = u_0'(x) = 0; \qquad u_0(x) = 1 \tag{43a,43b}$$

$$v_1(x) = u_1'(x) = -\frac{1}{Q} \tag{44a}$$

$$u_1(x) = -\frac{x}{Q} \tag{44b}$$

The second order equations are:

$$u_2'''' + Q(xu_2')' = -u_1(x) = \frac{x}{Q} \tag{45a}$$

$$u_2''(0) = 0, \quad u_2''(1) = 0, \quad u_2'''(0) = 0 \tag{45b,45c,45d}$$

$$u_2'''(1) + K_1 = 0 \tag{45e}$$

In terms of v_2, the equations become

$$v_2'' + Qxv_2 = \frac{x^2}{2Q} \tag{46a}$$

$$v_2'(0) = 0, \quad v_2'(1) = 0 \tag{46b,46c}$$

$$v_2(1) = \frac{1}{2Q^2} + \frac{K_1}{Q} \tag{46d}$$

where

$$v_2(x) = u_2'(x)$$

To solve Eqs. (46) let

$$v_2(x) = w(x) + \frac{x}{2Q^2}$$

In terms of $w(x)$, one has

$$w'' + Qxw = C \tag{47a}$$

$$w'(0) = -\frac{1}{2Q^2} \qquad w'(1) = -\frac{1}{2Q^2} \tag{47b,47c}$$

$$w(1) = \frac{K_1}{Q} \tag{47d}$$

Let $f(x)$ and $g(x)$ be the independent solutions to Eq. (47a) - Airy functions [12] - with boundary conditions $f(0) = 1$, $f'(0) = 0$ and $g(0) = 0$, $g'(0) = 1$ respectively. One then can write the solution to Eqs. (47) as:

$$w(x) = -\frac{1}{2Q^2}\left[\frac{1-g'(1)}{f'(1)} f(x) + g(x)\right] \tag{48}$$

From Eqs. (47d) and (48), one has

$$w(1) = \frac{K_1}{Q} = -\frac{1}{2Q^2}\left[\frac{1-g'(1)}{f'(1)} f(1) + g(1)\right] \tag{49a}$$

or

$$K_1 = -\frac{1}{2Q}\left[\frac{1-g'(1)}{f'(1)} f(1) + g(1)\right] = Qw(1) \tag{49b}$$

It should be observed that, if $Q = \hat{Q}_j$ of Eqs. (9), f'(1) becomes zero due to the fact that $\hat{Q}_j$'s are defined as such. In addition, w(1) can never be zero. To prove this, Eq. (47a) is multiplied by w'(x) and integrated over the interval (0, 1). One then arrives at

$$[w(1)]^2 = \int_0^1 [w(x)]^2 dx \tag{50}$$

Thus w(1) = 0 would imply $w(x) \equiv 0$ which can not be satisfied by Eqs. (47), and thus a contradiction is reached.

The fact $w(1) \neq 0$ for all values of Q indicates that it can not change sign as Q varies except possibly at those $\hat{Q}_j$'s of Eqs. (9), for which w(1) becomes infinite as f'(1) becomes zero. One can similarly show that when f'(1) = 0, f(1) can not be zero and thus f"(1) can not be zero due to Eq. (47a). Consequently w(1) does change sign at $\hat{Q}_j$. Since w(1) is proportional to K_1 (Eq. 49b) for all values of Q ($Q \neq 0$), K_1 can not be zero and will change sign at the $\hat{Q}_j$'s of Eqs. (9) and nowhere else. Now rewrite Eqs. (40) and (42) as

$$\lambda^2 = -6QK_\theta \text{ , for Q near zero} \tag{51a}$$

$$\lambda^2 = \frac{K_\theta}{K_1} \text{ , for Q away from zero and } K_\theta \text{ near zero} \tag{51b}$$

It is clear that we have proved what was stated at the beginning of this section that K_θ and λ^2 approach zero together for all $Q \neq 0$ (Figures 1 and 2). Also from Eq. (51b), since K_1 approaches infinity as Q approaches $\hat{Q}_j$, K_θ can assume very large values for small λ^2 near $\hat{Q}_j$. This indicates that the eigenvalue curves for all values of K_θ converge to the point $\lambda^2 = 0$, $Q = \hat{Q}_j$ (Figure 1).

All the analytical findings in this section have been observed in previous numerical results as shown in Figure 1.

With w(x) given in Eq. (48), one can write the solution to Eqs. (1') for $Q \neq \hat{Q}_j$

$$u(x) = 1 - \frac{\lambda^2}{Q} x - \frac{\lambda^4}{2Q^2} \left[\frac{1-g'(1)}{f'(1)} \int_0^x f(\xi)d\xi + \int_0^x g(\xi)d\xi - \frac{x^2}{2} \right] \tag{52a}$$

$$K_\theta = \lambda^2 K_1 \tag{52b}$$

and K_1 is to be evaluated from Eq. (49b). When Q is near $\hat{Q}_j$, the mode shape formula in Eq. (23) can be used, and the relationship

$$Q = Q_0 + \lambda^2 Q_1$$

applies. If Q is close to $\hat{Q}_j$ and K_θ is close to zero, Eqs. (23) and (52) should yield approximately the same results. In either formula, once two of the three parameters (K_θ, Q and λ^2) are chosen, the third one and the mode follow directly.

5. REPEATED EIGENVALUES WITH IDENTICAL EIGENFUNCTIONS. In Section 3 of this paper, we have observed two situations in which double zero eigenvalues exist with identical eigenfunctions.The first case is $K_\theta \neq -1$, $K_\theta \neq 0$ at those $\hat{Q}_j$'s of Eqs. (9); and the second case is $K_\theta = 0$ for all values of Q except $Q = 0$. The meaning of repeated eigenvalues with identical eigenfunctions will be explored further.

According to classical eigenvalue analysis [13], if only one independent eigenfunction can be found from Eqs. (1') (despite the presence of a repeated eigenvalue), then another linearly independent solution of the time equation, Eq. (1a), must exist in conjunction with that eigenfunction. This is analogous to the case of repeated roots in classical eigenvalue analysis. As an illustrative example, let us first consider the simple problem:

$$\begin{Bmatrix} \ddot{x}_1 \\ \ddot{x}_2 \end{Bmatrix} + \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} \begin{Bmatrix} x_1 \\ x_2 \end{Bmatrix} = \begin{Bmatrix} 0 \\ 0 \end{Bmatrix} \tag{53a}$$

and

$$\begin{bmatrix} \lambda^2 & 0 \\ 1 & \lambda^2 \end{bmatrix} \begin{Bmatrix} x_1 \\ x_2 \end{Bmatrix} = \begin{Bmatrix} 0 \\ 0 \end{Bmatrix} \tag{53b}$$

The eigenvalue problem in Eq. (53b) has repeated zero eigenvalue but only one eigenvector

$$\begin{Bmatrix} x_1 \\ x_2 \end{Bmatrix} = \begin{Bmatrix} 0 \\ 1 \end{Bmatrix} \tag{54}$$

The original differential equation (53a) however, has another solution independent of the one in Eq. (54)

$$\begin{Bmatrix} x_1 \\ x_2 \end{Bmatrix} = \begin{Bmatrix} 1 \\ 0 \end{Bmatrix} - \frac{t^2}{2} \begin{Bmatrix} 0 \\ 1 \end{Bmatrix} \tag{55}$$

One thus observes that while the orthogonal vector

$$\begin{Bmatrix} x_1 \\ x_2 \end{Bmatrix} = \begin{Bmatrix} 1 \\ 0 \end{Bmatrix}$$

can not be a solution to Eq. (53a) by itself, it can be one while driving the other eigenvector parametrically.

A similar situation exists in the present stability problem. To see this, let us consider a possible solution to Eqs. (1a) such that

$$u(x,t) = \tilde{u}\ (x) + \frac{u(0)}{2}\ t^2 \tag{56}$$

where u(0) is an arbitrary constant. Substitution of Eq. (56) into Eqs. (1) gives a set of equations for $\tilde{u}(x)$ which are identical to Eqs. (20). The solution is, therefore,

$$\tilde{u}(x) = u(0) \left[\frac{K_\theta}{(1+K_\theta)Q} \int_0^x \phi(\xi)d\xi - \frac{1}{Q}\ x \right] \tag{57}$$

For $K_\theta = 0$, Q can assume any non-zero value, and for $K_\theta \neq 0$, Q must be one of the $\hat{Q}_j$'s of Eqs. (9). No solution exists, however, for Q = 0 or for $K_\theta = -1$. It is recalled that $\tilde{u}(x)$ of Eq. (57) can not be a solution to Eqs. (1'). Thus the solution in the form of Eq. (56) exists just for those cases of repeated eigenvalues with identical eigenfunctions in the present stability problem. For $K_\theta = 0$, $\tilde{u}(x)$ of Eq. (57) is a rigid body rotation, and for $K_\theta \neq 0$, $\tilde{u}(x)$ is a combination of an integral of the Airy function and a rotation. Thus Beal [4] was correct in his intuitive prediction of the missing "modes". His only oversight was in calling them modes rather than Jordan functions.

The fact that $\tilde{u}(x)$ here is equal to $u_1(x)$ in Section 3 is more than coincidence. If we consider the limit as λ^2 approaches but is not identically equal to zero, then $\tilde{u}(x)$ does not come into picture. There are simply two independent eigenfunctions or mode shapes: $u(x) = u(0)$ and $u(x) = u_0(0) + \lambda^2 u_1(x)$ of Eqs. (23) or (33). The entire set of modes, including these two, form a complete set of functions. When $\lambda^2 \equiv 0$, however, the two modes in the neighborhood of $\lambda^2 = 0$ become identical and the function $u_1(x)$ is lost as an independent function. Therefore, $\tilde{u}(x)$ must equal $u_1(x)$ in order to form a complete set. However, $\tilde{u}(x)$ can not satisfy Eqs. (1'). Thus it takes on the role of a Jordon vector [14] that can parametrically excite the rigid body translation mode.

6. FURTHER NUMERICAL VERIFICATION. In the previous discussion, we have mentioned analytical proofs on several numerical findings as reported in Figures 1 and 2. Here we shall provide further comparison between the numerical results of finite element-unconstrained variational formulations* and the analytic formulas of asymptotic expansions in the neighborhood of $\lambda^2 = 0$.

*This finite element formulation has been described in detail elsewhere (see, for example, [8]).

6.1. Eigenvalues and Mode Shapes Near $Q = 0$. The analytic expression for the eigenvalue and the associated eigenfunction in the neighborhood of $Q = 0$ (but $Q \neq 0$) for all K_θ are given by Eqs. (40) and (41)

$$\lambda^2 = -6K_\theta Q$$

$$u(x) = 1 + \left(\frac{6K_\theta}{1-2K_\theta}\right)x$$

The agreement in eigenvalues between this formula and the finite element solution is shown in Table I. A typical mode shape comparison for $K_\theta = \pm 1$ and $Q = 0.001\pi^2$ is shown in Table II. The comparison indicates that the first term asymptotic solution is slightly better for $K_\theta = -1.0$ than it is for $K_\theta = 1.0$.

At $Q = 0$, the finite element formulation yields two zero eigenvalues and two rigid body modes - rigid body translation and rigid body rotation. This, of course, has been observed from Eqs. (1') in Section 2 earlier.

TABLE I. COMPARISON OF EIGENVALUE CALCULATION NEAR Q = 0
(ASYMPTOTIC VS. FINITE ELEMENT SOLUTIONS)

Q	K_θ	λ^2, Eq. (40b)	λ^2, F.E.
$0.001 \times \pi^2$	1.0	-5.922×10^{-2}	-5.917×10^{-2}
	-1.0	5.922×10^{-2}	5.923×10^{-2}
	5.0	-2.96×10^{-1}	-2.95×10^{-1}
	-5.0	2.96×10^{-1}	2.97×10^{-1}
$0.01 \times \pi^2$	1.0	-5.92×10^{-1}	-5.87×10^{-1}
	-1.0	5.92×10^{-1}	5.93×10^{-1}
	5.0	-2.96	-2.88
	-5.0	2.96	3.03

TABLE II. COMPARISON OF MODE SHAPES NEAR Q = 0
(ASYMPTOTIC VS. FINITE ELEMENT SOLUTIONS)

$Q = 0.001\pi^2$, $K_\theta = \pm 1.0$

	$K_\theta = -1.0$		$K_\theta = 1.0$	
x	u(x) Eq. (41)	u(x) F.E.	u(x) Eq. (41)	u(x) F.E.
0	1.0000	1.0000	1.0000	1.0000
1/9	0.7778	0.7778	0.3333	0.3330
2/9	0.5556	0.5556	-0.3333	-0.3340
3/9	0.3333	0.3334	-1.0000	-1.0010
4/9	0.1111	0.1112	-1.6667	-1.6680
5/9	-0.1111	-0.1110	-2.3333	-2.3348
6/9	-0.3333	-0.3333	-3.0000	-3.0016
7/9	-0.5556	-0.5556	-3.6667	-3.6682
8/9	-0.7778	-0.7780	-4.3333	-4.3348
1	-1.0000	-1.0003	-5.0000	-5.0013

6.2. Mode Shapes for Zero Eigenvalue at $Q = \hat{Q}_j$. For $K_\theta = -1$, the analytic solution in Section 2 states that $\lambda^2 = 0$ and the associated eigenfunction, other than the rigid body translation, is given by Eq. (12b):

$$u(x) = \int_0^x \phi(\xi)d\xi$$

where $\phi(x)$ has been defined in Eq. (10b). This expression of u(x) has been evaluated at $Q = \hat{Q}_1 = 2.598\pi^2$. This is compared with the finite element solution in Table III. The finite element eigenvalue solution associated with this mode shape is undistinguishable from the other known zero eigenvalue in magnitude.

Results from finite element computations also show that $K_\theta = -1$ is the only case where a bending mode, Eq. (12b), exists for $Q = \hat{Q}_j$ and for $\lambda^2 = 0$. For any other value of $K_\theta \neq -1.0$, only rigid body translation modes have been obtained. This agrees precisely with the analytical conclusions observed in Sections 2 and 3.

TABLE III. COMPARISON OF MODE SHAPE CALCULATIONS

$K_\theta = -1.0; \quad Q = \hat{Q}_1 = 2.598\pi^2; \quad \lambda^2 = 0$

x	u(x), Eq. (12b)	u(x), Eq. (12b) Normalized	u(x), F.E. Normalized
0	0.0	0.0	0.
1/9	-0.159077	0.325383	0.325337
2/9	-0.314905	0.644122	0.644035
3/9	-0.459358	0.939593	0.939462
4/9	-0.579979	1.186317	1.186158
5/9	-0.662340	1.354782	1.354613
6/9	-0.693969	1.419477	1.419323
7/9	-0.669319	1.369057	1.368941
8/9	-0.594394	1.215801	1.215744
1	-0.488890	1.000000	1.000000

6.3. Eigenvalue and Mode Shapes Near $Q = \hat{Q}_j$. The asymptotic solutions in this case are given by Eqs. (23) and (24)

$$u(x) = 1 + \frac{\lambda^2}{\hat{Q}_j}\left[\frac{K_\theta}{1+K_\theta}\int_0^x \phi(\xi)d\xi - x\right]$$

$$Q = \hat{Q}_j + \lambda^2 Q_1$$

where $K_\theta \neq -1$ and Q_1 is evaluated by Eq. (32a). This eigenvalue λ^2 and mode shape u(x) in the expressions above have been evaluated for several values of Q and K_θ and they are compared with the finite element solutions in Tables IV and V.

TABLE IV. COMPARISON OF EIGENVALUE CALCULATIONS (ASYMPTOTIC VS. FINITE ELEMENT SOLUTIONS)

K_θ	$Q = \hat{Q}_1 + 0.01\pi^2$		$Q = \hat{Q}_1 - 0.01\pi^2$	
	λ^2, Eq. (24)	λ^2, F.E.	λ^2, Eq. (24)	λ^2, F.E.
0.01	0.00680	0.00687	-0.00680	-0.00667
-0.01	-0.00706	-0.00714	0.00706	0.00692
0.1	0.05839	0.05928	-0.05839	-0.05753
-0.1	-0.08521	-0.08669	0.08521	0.08378
1.0	0.24165	0.24477	-0.24165	-0.23863

TABLE V. COMPARISON OF MODE SHAPE CALCULATIONS (ASYMPTOTIC VS. FINITE ELEMENT SOLUTIONS)

	$Q = \hat{Q}_1 - 0.01\pi^2$, $K_\theta = 1.0$		$Q = \hat{Q}_1 + 0.01\pi^2$, $K_\theta = 1.0$	
x	Asymptotic $\lambda^2 = -0.2416$, Eq. (24) u(x), Eq. (23)	F.E. $\lambda^2 = -0.2386$ u(x)	Asymptotic $\lambda^2 = 0.2416$, Eq. (24) u(x), Eq. (23)	F.E. $\lambda^2 = 0.2448$ u(x)
0	1.000000	1.000000	1.000000	1.000000
1/9	1.001836	1.001789	0.998165	0.998195
2/9	1.003655	1.003563	0.996345	0.996405
3/9	1.005418	1.005283	0.994582	0.994669
4/9	1.007063	1.006891	0.992937	0.993046
5/9	1.008518	1.008320	0.991482	0.991604
6/9	1.009722	1.009512	0.990278	0.990403
7/9	1.010647	1.010440	0.989353	0.989468
8/9	1.011323	1.011132	0.988677	0.988772
1	1.011847	1.011681	0.988153	0.988221

The mode shapes in Table V show that the rigid body translation represents the major portion of the motion. However, both rigid body rotation and bending are present.

6.4. Zero Curvature at $\lambda^2 = 0$ and $Q = \hat{Q}_j$. It was noted in Section 3 that the curvature of a "Q vs. λ" curve (Figure 1 or 2 rotated 90°) at $\lambda = 0$ and $Q = \hat{Q}_j$ is

$$\rho = 2Q_1$$

From Eq. (32a) it is clear that ρ is a function of K_θ. Since a positive ρ implies that the curve is "concave" in the lower half plane and "convex" in the upper half plane, and a negative ρ implies the reverse (Figure 3), it is of interest to find out at what value of K_θ this change-over does take place. Letting $\rho = 2Q_1 = 0$, Eq. (32a) yields

$$K_\theta = -0.53543$$

It would not be convenient to evaluate this K_θ by the finite element method. However, the curves as shown in Figure 2 appears to agree with the calculation given above.

6.5. Eigenvalues and Mode Shapes for K_θ Near Zero. For K_θ in the neighborhood of zero, the asymptotic solutions are given by Eqs. (52) as

$$u(x) = 1 - \frac{\lambda^2}{Q} x$$

$$K_\theta = \lambda^2 K_1$$

for all values of Q ($Q \neq 0$), where K_1 can be calculated by Eq. (49b). Results from the above expression and the finite element solutions are compared for several values of Q and K_θ as shown in Table VI.

As we have noted near the end of Section 4, Eqs. (23), (24) and Eqs. (52) should yield approximately same results for K_θ near zero and for Q near $\hat{Q}_j$. This fact is demonstrated by the numerical results given in Table VII.

7. CONCLUSIONS. From the analysis presented in this paper one can clearly conclude the following:

1. A double zero eigenvalue, $\lambda^2 = 0$, exists at $Q = 0$ for all values of K_θ. There exist two independent eigenfunctions: a rigid body translation and a rigid body rotation, for the repeated roots.

2. For $K_\theta = -1$, double zero eigenvalues exist only at a discrete set of $\hat{Q}_j$'s, $i = 0,1,2,\ldots$, defined in Eqs. (9). At each $\hat{Q}_j$ there exist two independent eigenfunctions: a rigid body translation and an integrated Airy function. For $Q_0 = 0$, this integrated Airy function reduces to a rigid body rotation.

3. For $K_\theta = 0$, double zero eigenvalues exist for all values of Q and the only eigenfunction for these repeated eigenvalues (except at $Q = 0$) is a rigid body translation.

4. For K_θ near zero, but not identically zero, one can write $K_\theta = \lambda^2 K_1$ or $\lambda^2 = K_\theta / K_1$ where K_1 is never zero for any values of Q and becomes infinite when $Q = \hat{Q}_j$ of Eqs. (9). The eigenfunction associated with this λ^2 is a combination of a unit rigid body translation, a rigid body rotation proportioned to λ^2, and other bending terms of still higher order in λ^2.

5. For $K_\theta \neq 0$ and $K_\theta \neq -1$, double zero eigenvalues exist only at those $\hat{Q}_j$'s of Eqs. (9) as for the case $K_\theta = -1$. However, unlike the case $K_\theta = -1$, only one eigenfunction exists at these $\hat{Q}_j$, and it is a rigid body translation.

TABLE VI. COMPARISON OF EIGENVALUE AND MODE CALCULATIONS
K_θ NEAR ZERO, Q NOT NEAR $\hat{Q}_j$

	$K_\theta = 0.01$, $Q = \pi^2$		$K_\theta = -0.01$, $Q = \pi^2$		$K_\theta = 0.1$, $Q = \pi^2$	
x	Asymptotic $\lambda^2 = -0.3859$ Eq. (52b) u(x), Eq. (52a)	F.E. $\lambda^2 = -0.3837$ u(x)	Asymptotic $\lambda^2 = 0.3859$ Eq. (52b) u(x), Eq. (52a)	F.E. $\lambda^2 = 0.3882$ u(x)	Asymptotic $\lambda^2 = -3.8592$ Eq. (52b) u(x), Eq. (52a)	F.E. $\lambda^2 = -3.6470$ u(x)
0	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000
1/9	1.00435	1.00440	0.99566	0.99571	1.04345	1.04945
2/9	1.00869	1.00880	0.99131	0.99141	1.08689	1.09885
3/9	1.01303	1.01319	0.98697	0.98712	1.13034	1.14805
4/9	1.01738	1.01759	0.98262	0.98283	1.17379	1.19687
5/9	1.02172	1.02198	0.97828	0.97852	1.21723	1.24513
6/9	1.02607	1.02636	0.97393	0.97421	1.26068	1.29265
7/9	1.03041	1.03073	0.96959	0.96990	1.30413	1.33938
8/9	1.03476	1.03510	0.96524	0.96557	1.34757	1.38541
1	1.03910	1.03946	0.96090	0.96124	1.39102	1.43100

TABLE VII. COMPARISON OF EIGENVALUE AND MODE CALCULATIONS K_θ NEAR ZERO AND Q NEAR $\hat{Q}_j$

	$K_\theta = -0.01$, $Q = Q_1 + 0.01\pi^2$			$K_\theta = 0.1$, $Q = \hat{Q}_1 - 0.1\pi^2$		
	Asymptotic		F.E.	Asymptotic		F.E.
x	$\lambda^2 = -0.00716$, Eq. (24) u(x), Eq. (23)	$\lambda^2 = -0.00718$, Eq. (52b) u(x), Eq. (52a)	$\lambda^2 = -0.00714$ u(x)	$\lambda^2 = -0.05839$, Eq. (24) u(x), Eq. (23)	$\lambda^2 = -0.06688$, Eq. (52b) u(x), Eq. (52a)	$\lambda^2 = -0.05753$ u(x)
0	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000
1/9	1.000030	1.000031	1.000030	1.000288	1.000291	1.000283
2/9	1.000060	1.000062	1.000061	1.000575	1.000582	1.000565
3/9	1.000090	1.000093	1.000091	1.000859	1.000873	1.000845
4/9	1.000121	1.000124	1.000122	1.001138	1.001164	1.001120
5/9	1.000151	1.000155	1.000152	1.001409	1.001455	1.001387
6/9	1.000182	1.000186	1.000183	1.001669	1.001745	1.001644
7/9	1.000212	1.000217	1.000214	1.001917	1.002036	1.001890
8/9	1.000243	1.000248	1.000245	1.002154	1.002327	1.002125
1	1.000274	1.000279	1.000276	1.002384	1.002618	1.002354

6. For $K_\theta \neq 0$, $K_\theta \neq -1$ and for λ^2 near zero but not identically zero, the eigenfunction associated with λ^2 near the $\hat{Q}_j$'s of Eqs. (9) is a combination of a unit rigid body translation, a rigid body rotation, and a bending of integrated Airy function. The latter two terms are proportional to λ^2.

7. For $\hat{Q}_j < Q < \hat{Q}_{j+1}$, where $\hat{Q}_j$'s are defined in Eqs. (9), λ^2 is negative (stable vibrations) for a positive K_θ and λ^2 is positive (divergence instability) for a negative K_θ if $j = 0,2,4.....$ The sign of λ^2 (and thus the stability characteristics) is reversed in the above statement if $j = 1, 3,5,...$

8. For either $K_\theta = 0$ and all values of Q, or $K_\theta \neq 0$, $K_\theta \neq -1$ and $Q = \hat{Q}_j$ of Eqs. (9), the solution of Eqs. (1') has repeated zero eigenvalues with identical rigid body translation mode. Another independent solution of Eqs. (1) can be obtained by the method of variation of parameters as a Jordon vector in the form Eqs. (56) and (57).

ACKNOWLEDGEMENTS. The authors wish to express their appreciation to Messrs. R. F. Haggerty and R. W. Soanes of Benet Weapons Laboratory for their assistance in obtaining the numerical data.

REFERENCES.

1. S. Timoshenko and D. H. Young, 1955 Vibration Problems in Engineering, Van Nostrand, Princeton, p. 374.

2. Ibid, p. 336.

3. S. Silverberg, 1959 Space Technology Laboratories, Inc. Technical Report TR-59-0000-00791. The Effect of Longitudinal Acceleration Upon the Natural Modes of Vibration of a Beam.

4. T. R. Beal, 1965 American Institute of Aeronautics and Astronautics Journal, 3, 486-494. Dynamic Stability of a Flexible Missile Under Constant and Pulsating Thrust.

5. V. I. Feodos'ev, 1965 Prikladnaia Matematika I Mekhanika, 29, 391-392. On a Stability Problem (translated from Russian).

6. G. Y. Matsumoto and C. D. Mote, 1972 Journal of Dynamic Systems, Measurement and Control, Transaction of American Society of Mechanical Engineers, 94, 330-334. Time Delay Instability in Large Order Systems with Controlled Follower Force.

7. J. J. Wu, 1975 Journal of Sound and Vibration, 43, 45-52. On the Stability of a Free-Free Beam Under Axial Thrust Subjected to Directional Control.

8. J. J. Wu, 1976 Journal of Sound and Vibration, 46, 51-57. On Mode Shapes of a Stability Problem.

9. V. Sundararamaiah and D. J. Johns, 48, pp. 571-574. Comments on "On the Stability of a Free-Free Beam Under Axial Thrust Subjected to Directional Control".

10. J. J. Wu, 1976 Journal of Sound and Vibration, 49 (1), pp. 141-147. On Missile Stability.

11. V. Sundararamaiah and D. J. Johns. Private Communication.

12. M. Abramowitz and I. A. Stegun, Editors, Handbook of Mathematical Functions, Dover, NY, 1970, pp. 446-452.

13. J. H. Wilkonson, The Algebraic Eigenvalue Problem, Oxford Press, 1965, pp. 9-11, 31-32.

14. Ibid, pp. 523-540.

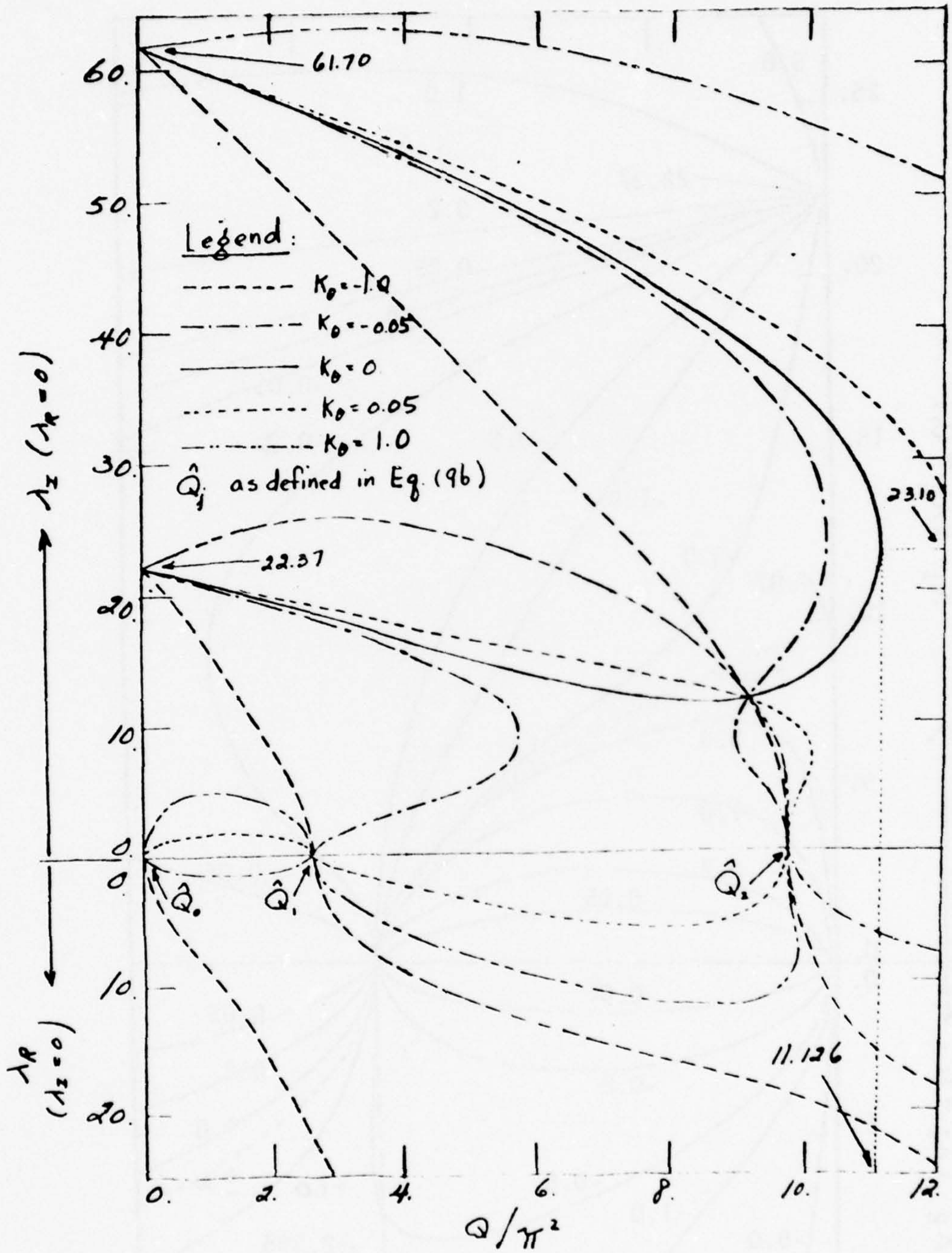


FIGURE 1. Curves of Eigenvalue λ (λ_I and λ_R) vs. Thrust Q for Various Values of K_θ (Finite Element Results).

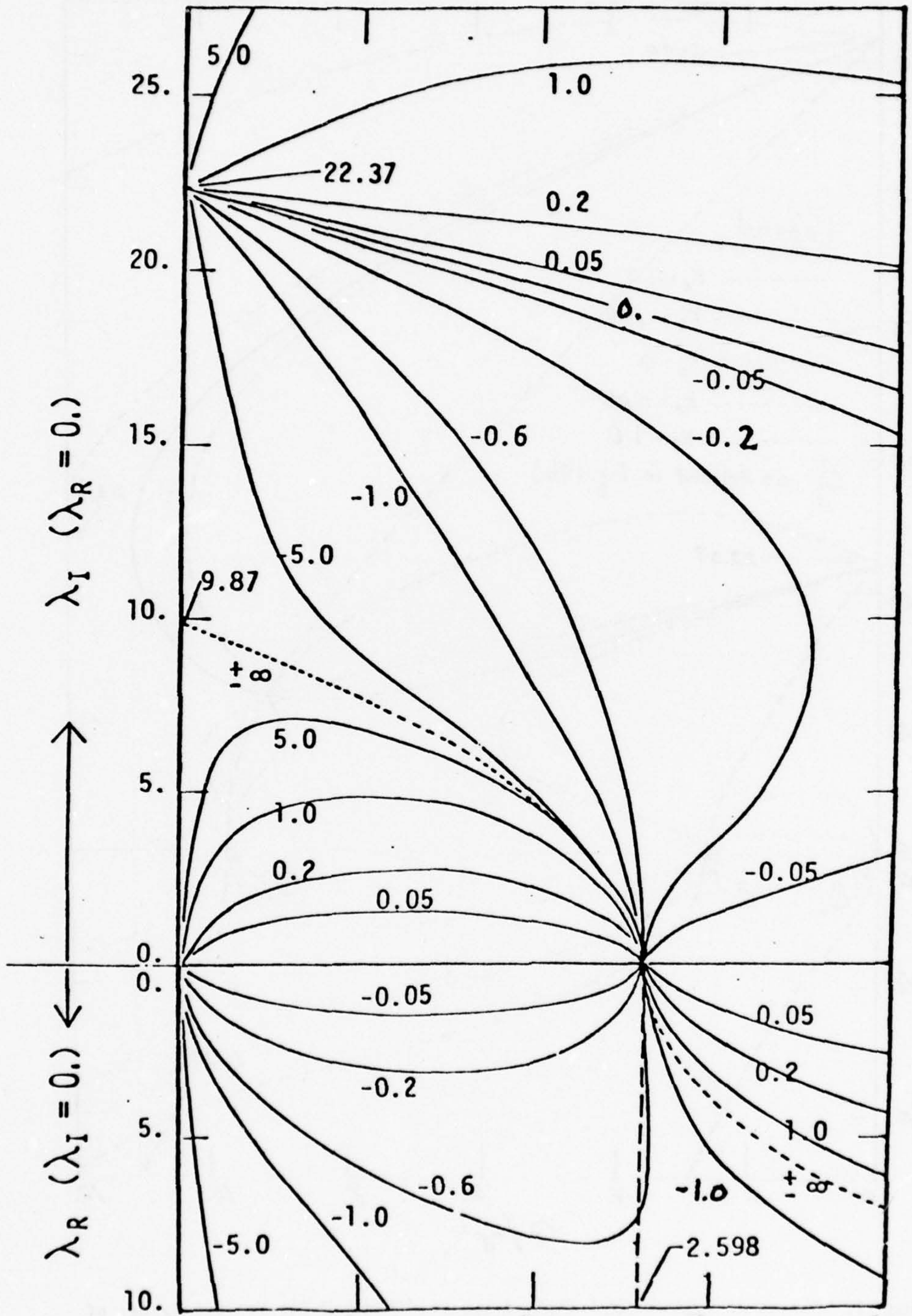


FIGURE 2. Curves of Eigenvalue λ (λ_I and λ_R) vs. Thrust Q for Various Values of K_θ in a Larger Scale (Finite Element Results, Reproduced from Reference [8]).

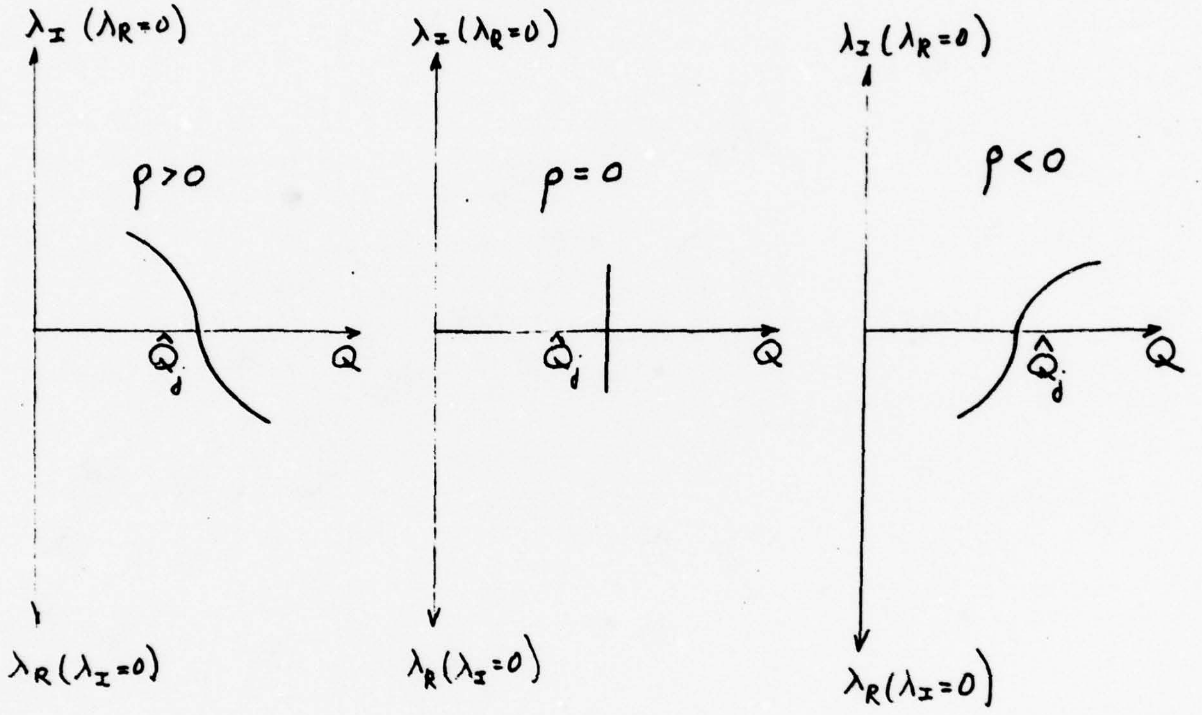

$\lambda_I (\lambda_R = 0)$
$\rho > 0$
$\hat{Q}_j$
Q
$\lambda_R (\lambda_I = 0)$
$\lambda_I (\lambda_R = 0)$
$\rho = 0$
$\hat{Q}_j$
Q
$\lambda_R (\lambda_I = 0)$
$\lambda_I (\lambda_R = 0)$
$\rho < 0$
$\hat{Q}_j$
Q
$\lambda_R (\lambda_I = 0)$

FIGURE 3. Curvature Characteristics of a Q vs. λ Curve at $Q = \hat{Q}_j$ and $\lambda = 0$.

AN INVERSE HYPERBOLIC BOUNDARY VALUE PROBLEM*

William W. Symes
Mathematics Research Center
University of Wisconsin-Madison
Madison, WI 53706

ABSTRACT. We solve an inverse boundary value problem for a two-dimensional wave equation. This problem is prototypical of a number of inverse problems of applied mathematics - for instance, the main problem of seismology has the same general nature. Our solution proceeds via a nonlinear Volterra equation, and is constructive, with explicit error bounds.

I. INTRODUCTION.

The general setting of the problem treated here is the continuum mechanics of wave propagation. We consider a continuous medium confined to a region Ω with boundary $\partial\Omega$, in which disturbances propagate at finite speed - hence, according to a hyperbolic system of partial differential equations. We treat small disturbances only, hence assume that the system is linear.

The quintessential example is the system of equations of linear elasticity theory:

$$\frac{\partial^2 \underline{u}}{\partial t^2}(x,t) = \frac{1}{\rho(x)} \nabla \cdot \underline{\underline{\tau}}(\lambda(x),\mu(x),\underline{u}(x,t)), \qquad x \in \Omega \subset \mathbb{R}^3 . \tag{1}$$

Here $\underline{u}(x,t) = (u_1(x,t),u_2(x,t),u_3(x,t))$ is the infinitesimal displacement vector, $\underline{\underline{\tau}}$ is the stress tensor, $\rho(x)$ is the density, and $\lambda(x)$, $\mu(x)$ are the Lamé "constants" which characterize the mechanical properties of the medium at the location $x \in \Omega$. Usually, boundary conditions are imposed on $\partial\Omega$, of the form

$$\underline{\underline{\tau}}(\lambda(x),\mu(x),u(x,t)) \cdot \underline{n}(x) \equiv 0, \qquad \forall x \in \partial\Omega \tag{2}$$

where $\underline{n}(x)$ is the normal vector at $x \in \partial\Omega$.

Another, much simpler example is the two dimensional wave equation with potential on the half-axis

$$\frac{\partial^2 u}{\partial t^2}(x,t) = \frac{\partial^2 u}{\partial x^2}(x,t) + q(x)u(x,t), \qquad x \geq 0 , \tag{3}$$

*A more detailed treatment of this work can be found in References (1), (2).

Sponsored by the United States Army under Contract No. DAAG29-75-C-0024 and the National Science Foundation under Grant No. MCS75-17385 A01.

with "energy-preserving" (self-adjoint) boundary condition of the form

$$\frac{\partial u}{\partial x}(0,t) + hu(0,t) \equiv 0 \tag{4}$$

where h is some real number (or ∞, in which case the boundary condition reads $u(0,t) \equiv 0$). This system models a number of physical processes - for instance, certain kinds of lossless transmission lines. Our main interest, however, is in using this simple system as a mathematical prototype of more complicated systems, such as linear elasticity.

The sort of problem we wish to solve is the inverse boundary value problem. The unknowns of our problem are the coefficients of the hyperbolic equations of motion, and perhaps the boundary conditions. The data of our problem are the motions of the system on the boundary, resulting from a known applied force.

In the linear elasticity example, we are thus given an external force $\underline{F}(x,t)$, $x \in \Omega$, $t \geq 0$, and observations of the displacement vector on the boundary: $\{\underline{u}(x,t), x \in \partial\Omega, t \geq 0\}$. We know that $\underline{u}$ solves the boundary value problem

$$\ddot{\underline{u}} = \frac{1}{\rho} \nabla \cdot \underline{\underline{\tau}} + \underline{F}, \quad x \in \Omega, \quad t \geq 0\ ,$$

$$\underline{\underline{\tau}} \cdot \underline{n} = 0 \ \text{ on } \ \partial\Omega, \quad \underline{u}(x,0) \equiv 0, \quad x \in \Omega\ .$$

We are to deduce the density ρ and the Lamé constants λ and μ in the region Ω.

In this form, the inverse boundary value problem for linear elasticity is recognizable as the main problem of seismology, in an idealized version. Experience in seismology suggests the heuristic rule: in order to extract as much information as possible about the system in the inaccessible interior of the region Ω, the applied force $\underline{F}$ should be as sharply localized in time and space as possible, i.e. approximate a δ-function (unit impulse), in order to excite all modes of the system.

With these ideas in mind, we make up a simpler model problem. We consider the two-dimensional problem (3), (4), and ask whether the potential $q(x)$, $x \geq 0$, and the boundary condition parameter h can be recovered from the response of the system $\{u(0,t) : t \geq 0\}$ on the boundary to a unit impulse concentrated at $x = t = 0$. Actually, it is more convenient to transform this problem, via a standard trick (Duhamel's integral), into the form:

Let $u(x,t)$, $x,t \geq 0$, be the solution of the initial-boundary value problem

$$\begin{aligned} &u_{tt} - u_{xx} + qu = 0\ , \\ &u_x(0,t) + hu(0,t) \equiv 0\ , \\ &u(x,0) = \delta(x)\ , \\ &u_t(x,0) = 0\ . \end{aligned} \tag{5}$$

Set $\tilde{f}(t) = u(0,t)$, $t > 0$. The well-known theory of the Cauchy problem (see e.g. reference (3)) asserts the existence of a unique solution u, hence a unique $\tilde{f}$, given the coefficients $q(x)$ and h. Let S denote this correspondence:

$$S : \{h, q(x), x \geq 0\} \to \{\tilde{f}(t), t \geq 0\} .$$

Our problem may now be phrased: to determine whether the correspondence S is invertible. Other related questions are:

(1) The question of extent: what functions $\tilde{f}$ can appear as boundary values of the solution of a problem of the form (5), as h, $q(x)$ range over all possible choices?

(2) The question of stability: is the inverse of the correspondence S continuous in some sense? I.e., can we guarantee that small errors in the measurement of $\tilde{f}$ do not result in huge uncontrollable errors in the recovery of h, $q(x)$?

We are able to answer all of these questions. Moreover, we are able to give a constructive procedure (i.e., one involving only algebraic manipulations and quadratures) for the recovery of h, $q(x)$ from $\tilde{f}(t)$ - with explicit, best-possible error estimates.

II. STATEMENT OF RESULTS. The precise statement of our results requires that we specify the class of permissible coefficient functions $q(x)$. The correct choice turns out to be the Sobolev space $W^{m,1}_{loc}(\mathbb{R}^+)$ (here, and in the following, we use the notation $\mathbb{R}^+ = [0,\infty) \subset \mathbb{R}$ for the half-axis). This space consists of all $m - 1$ - times continuously differentiable functions whose m^{th} derivatives, moreover, though not necessarily continuous, exist as locally absolutely integrable functions. Thus, for example, $W^{0,1}_{loc}$ contains functions which have continuous derivatives except at various isolated points, where they may have "corners".

To discuss stability, we need some notion of closeness, or approach to zero. A sequence $\{f_n\}$ of functions in $W^{m,1}_{loc}(\mathbb{R}^+)$ is said to approach zero iff for all $T > 0$, the sequences of numbers

$$\sup_{x\in[0,T]} \left|\frac{d^k}{dx^k} f_n(x)\right|, \qquad k = 0,\ldots,m - 1 ,$$

and

$$\int_0^T \left|\frac{d^m f_n}{dx^m}(x)\right| dx$$

approach zero as $n \to \infty$. The question of extent is answered by

Theorem I. A function $\tilde{f}(t)$, $t \geq 0$, is the boundary value of the solution to problem (5) for some choice $q \in W^{m,1}_{loc}(\mathbb{R}^+)$ and $h \in \mathbb{R}$, if and only if:

1) $\tilde{f} \in W^{m+1,1}_{loc}(\mathbb{R}^+)$

2) $\tilde{f}(0) = h$

3) The kernel $f(s,t) = \frac{1}{2}(\tilde{f}(s+t) + \tilde{f}(s-t))$ satisfies the condition: for any $T > 0$ there exists $\varepsilon(T) > 0$ so that for any $u \in L^2[0,T]$,

$$\int_0^T |u|^2 + \int_0^T\int_0^T dsdt\ u(s)\bar{u}(t)f(s,t) \geq \varepsilon(T) \int_0^T |u|^2$$

(in interpreting 3, $\tilde{f}$ is extended to be an even function: $\tilde{f}(-t) = \tilde{f}(t)$).

The question of the invertibility of S, and the stability question, are answered by

Theorem II.

1) The set of $\tilde{f} \in W^{m+1,1}_{loc}(\mathbb{R}^+)$ defined by condition 3) of Theorem I forms an open cone $C \subset W^{m+1,1}_{loc}(\mathbb{R}^+)$.

2) The correspondence $S : (h,q) \in \mathbb{R} \times W^{m,1}_{loc}(\mathbb{R}^+) \to \tilde{f} \in C$ is invertible:

3) The inverse correspondence $\tilde{f} \to (h,q)$ is continuous as a mapping $W^{m+1,1}_{loc}(\mathbb{R}^+) \supset C \to \mathbb{R} \times W^{m,1}_{loc}(\mathbb{R}^+)$.

Moreover, as mentioned above, the continuity statement 3) of Theorem II is effected by a collection of error estimates. The exact form of these error estimates is detailed in reference (1). Their essence is that the relative amount of error tends to increase as the data $\tilde{f}$ is chosen closer to the boundary of the cone C.

III. SKETCH OF PROOF. The solution $u(x,t)$ of (5) is a part of the Riemann function for the boundary value problem

$$\begin{aligned} &u_{tt} - u_{xx} + qu = 0, \qquad t,x \geq 0, \\ &u_x(0,t) + hu(0,t) \equiv 0\ . \end{aligned} \tag{6}$$

The Riemann Function $R(x,t,x_0,t_0)$ is the solution of (6) with the initial conditions

$$R(x,t_0,x_0,t_0) = \delta(x - x_0)$$

$$R_t(x,t_0,x_0,t_0) = 0\ .$$

Thus $u(x,t) = R(x,t,0,0)$.

The results outlined in Section II proceed from three facts regarding the Riemann function:

1) (Group Property) Let $U(t)$ be the operator which maps Cauchy Data $(u(x,t_0,u_t(x,t_0))$ for a solution of (6) at time t_0 to the Cauchy data for the same solution $(u(x,t+t_0),u_t(x,t+t_0))$ at time $t+t_0$. Then U is independent of t_0, and

$$U(s)U(t) = U(s+t) \; .$$

2) The Riemann Function propagates Cauchy data for (6). That is, if $u(x,t)$ is an arbitrary solution of (6) then

$$\begin{pmatrix} u(x,t) \\ u_t(x,t) \end{pmatrix} = \int dx_0 \; \mathcal{R}(x,t,x_0,t_0) \begin{pmatrix} u(x_0,t_0) \\ u_t(x_0,t_0) \end{pmatrix} .$$

where $\mathcal{R}$ is the 2×2 matrix

$$\mathcal{R}(x,t,x_0,t_0) = \begin{pmatrix} R(x,t,x_0,t_0) & \int_{t_0}^{t} d\sigma R(x,t,x_0,\sigma) \\ \frac{\partial}{\partial t} R(x,t,x_0,t_0) & \frac{\partial}{\partial t} \int_{t_0}^{t} d\sigma R(x,t,x_0,\sigma) \end{pmatrix} .$$

Thus $\mathcal{R}$ is the kernel which implements the operator U.

3) (Progressing Wave Expansion). The distribution $u(x,t) = R(x,t,0,0)$ can be decomposed:

$$u(x,t) = \delta(x+t) + \delta(x-t) + K(x,t)$$

where $K(x,t) = K(x,-t)$ has one more derivative than q in the region $0 \leq x \leq |t|$, vanishes outside this region, and on the boundary satisfies the <u>transport</u> <u>equation</u>:

$$K(x,x) = h - \frac{1}{2}\int_0^x q \; . \tag{7}$$

Properties 1) and 2) imply an integral equation for $\mathcal{R}$, from which an integral equation for R may be extracted. A special case is the equation

$$(\delta(x+t) + \delta(x-t)) + f(s,t) = \int_0^\infty dy \; u(y,s)u(y,t)$$

where, as before, $f(s,t) = \frac{1}{2}(\tilde{f}(s+t) + \tilde{f}(s-t))$, $\tilde{f}(t) = u(0,t)$, $t > 0$.

Combined with 3) this can be rewritten

$$f(s,t) = K(s,t) + \int_0^s dy\, K(y,s)K(y,t) \tag{8}$$

say for $s \leq t$. This nonlinear Volterra equation is solved by a combination of iteration and a priori estimates; the source of the estimates, and a necessary and sufficient condition that (8) have a solution, is condition 3) of Theorem I, i.e. that $f \in C$. The potential q is then recovered from the solution K of (8) by the transport equation (7); see (1) for details.

IV. CONCLUDING REMARKS. We conclude by describing related work by other authors.

This work began as an attempt to understand the paper of Gel'fand and Levitan on the inverse spectral problem for Sturm-Liouville operators, which is a time-independent version of the problem considered here. (Reference (4); see (1) for an explanation of the relation.) Indeed, the invertibility and extent parts of Theorems I and II, though not the stability statement, are implicitly contained in the sharp version due to Levitan and Gasymov (ref. 5) of the results of (4). The method of proof of these authors is considerably different, however. They derive equations (7) and (8), by different techniques, and then linearize equation (8), to obtain a Fredholm integral equation, which is then solved by the Fredholm Alternative Theorem. Our constructive solution to equation (8) provides an obvious contrast, and seems more suitable for numerical work. Most providentially, however, equation (8) has an analogue for inverse problems involving systems with several sound speeds, including some which more closely model linear elasticity; the Fredholm equation approach, on the other hand, does not work for multiple sound speeds at all. Some consequences of this fortunate circumstance, including analogues of Theorems I and II, will appear in reference (2). Finally, equations (7) and (8) also have obvious analogues for several-dimensional problems; there, however, the analytical details are much more difficult, and it remains to be seen whether our methods will suffice to give solutions to these problems.

REFERENCES

(1) Symes, W., Inverse Boundary Value Problems and a Theorem of Gel'fand and Levitan, MRC Technical Summary Report #1846, University of Wisconsin, Madison, April 1978. (To appear in J. of Math. Anal. and Appl.)

(2) Symes, W., Inverse Boundary Value Problems for Hyperbolic Systems, to appear.

(3) Courant, R., and Hilbert, D., Methods of Mathematical Physics, Vol. II, Wiley-Interscience, N.Y. 1962.

(4) Gel'fand, I. M., and Levitan, B. M., On the determination of a differential equation from its spectral function, AMS Tr. (2) 1, 253-304 (1955).

(5) Levitan, B. M., and Gasymov, M., Russ. Math. Surveys (1964).

SOME ANALYTICAL ASPECTS OF A NONLINEAR TRANSIENT ELECTROMAGNETIC FIELD PENETRATION PROBLEM

William J. Croisant and Paul Nielsen
U.S. Army Construction Engineering Research Laboratory
Champaign, Illinois 61820

ABSTRACT

The calculation of the transient electromagnetic field penetration into an electrically conducting medium is complicated considerably if the permeability is a function of magnetic field strength. Even in seemingly simple problems such calculations are difficult because the governing partial differential equations are nonlinear. This paper examines the penetration of a step increase in magnetic field strength into a semi-infinite medium that has a permeability which is a function of magnetic field strength. The problem can be simplified through a simple transformation of variables which reduces the number of independent variables by one. Although the above transformation can be performed regardless of the form of the permeability function, each form of the permeability results in a different second order nonlinear differential equation so that a particular permeability function must be specified to proceed with an analysis. Several permeability functions are considered in this paper. For a constant permeability the differential equation is linear and can be easily solved to obtain the classic solutions for the magnetic and electric field distributions. The perturbation of these solutions due to small variations in permeability is then derived. In addition, a permeability which is an exponentially decreasing function of magnetic field strength is examined. Although a formal parametric solution for this case exists, a simple closed form analytical solution in terms of elementary functions does not seem to exist; therefore, the analytical approach consists of mathematical analysis

suplemented with numerical calculations. The results are compared to the classic solution for the linear case with constant permeability.

INTRODUCTION

If a pulse of electric current is suddenly forced to flow along the surface of an electrically conducting medium, electromagnetic fields are induced which oppose the spread of this current. Hence, the current and the electromagnetic fields associated with it tend to spread, or diffuse, more or less slowly throughout the medium at a rate which depends on the permeability and the conductivity of the medium. The calculation of the transient distribution of such electromagnetic fields within a conducting medium is of interest for a number of areas (an important one being electromagnetic shielding) and has received a good deal of attention.

The partial differential equations which govern such electromagnetic field propagation are linear if the material properties of the medium (the conductivity and the permeability) are independent of the applied electromagnetic field level. The analytical theory for electromagnetic field calculations for such media has been well developed and extensive analyses exist for a variety of boundary conditions. Frequently, the exact analytical solution can be derived using any of a number of standard linear mathematical techniques.

The calculation of electromagnetic field penetration in ferromagnetic materials (e.g., the elements iron, nickel, and cobalt as well as certain alloys) is considerably more complicated because the permeabilities of these materials vary with applied magnetic field strength. For such materials the partial differential equations which govern electromagnetic field propagation are nonlinear and their analytic solution is considerably more complex. The analytical situation is not as satisfactory as in the field independent case because most of the standard linear mathematical techniques cannot be applied. Due to the difficulties associated with the nonlinear partial differential equations, relatively few analyses are available which can be used to assess the effect of a field-dependent permeability on a time-dependent electromagnetic field penetration problem.

Since there is no general analytical theory for nonlinear differential equations that is applicable to all forms of nonlinearity and all boundary conditions, each nonlinear problem is usually considered as a separate and distinct obstacle to be overcome. This study considers the time-dependent penetration of a step increase in magnetic field strength into a semi-infinite conducting medium which has a permeability that is a function of magnetic field strength. The medium is considered to be homogeneous, isotropic, and initially demagnetized. It is presumed that the medium at any point follows its initial magnetization curve. (The phenomenon of hysteresis does not arise in this problem.)

The problem described above can be simplified from one involving a second order nonlinear partial differential equation to one involving a second order nonlinear ordinary differential equation through a simple transformation of variables. This transformation can be employed regardless of the form of the differential permeability. This reduction in the number of independent variables is a significant simplification of the problem because ordinary differential equations are usually easier to deal with than partial differential equations whether by exact, approximate, or numerical analysis.

While this simplification does not depend on the functional form of the permeability, a representation of the differential permeability must be specified to proceed further. Each form for the differential permeability results in a second order differential equation possessing its own degree of perversity. The general approach is, therefore, to use a functional form for the permeability that adequately represents the variation of the permeability over some range of magnetic field strength and is such that properties of the solution can be determined.

Traditionally the permeability has been taken to be constant giving the classic solution to the linear case, and, for purposes of comparison, this study first considers this classic solution. Secondly, if the differential permeability exhibits only moderate change with magnetic field strength, then the change can be taken to be proportional to the magnetic field strength. This study examines the case of a permeability which exhibits small changes with increasing magnetic field strength. Perturbation analysis is used to determine a first order correction to the solution for the linear case. Finally, to examine the effects of a larger variation in variable permeability this study considers a permeability which decreases exponentially with increasing magnetic field strength. The analytical approach consists of mathematical analysis supplemented with numerical calculations. Although a formal parametric solution exists, a simple closed form solution in terms of elementary functions does not seem to exist. For this reason several analytical techniques, as well as numerical calculations, have been employed to deduce properties of the solution.

Numerical results for an exponentially decreasing permeability are presented. For small changes in the differential permeability, the solutions for the electromagnetic fields differ only slightly from the classic solution for the linear case. On the other hand, in cases of large changes in permeability the solutions to nonlinear cases are observed to differ markedly from the solution to the linear case not only in magnitude but also in shape.

MATHEMATICAL FORMULATION OF THE PROBLEM

The propagation of electromagnetic fields in a conducting medium (where the displacement current can be neglected) is governed by the equations

$$\nabla \times \vec{E} = -\frac{\partial \vec{B}}{\partial t} \tag{1}$$

and

$$\nabla \times \vec{H} = \sigma \vec{E} \tag{2}$$

where $\vec{E}$ is the electric field, $\vec{H}$ is the magnetic field strength, $\vec{B}$ is the magnetic flux density, and σ is the electrical conductivity. In a homogeneous and isotropic medium, it is assumed that $\vec{B} = B(\vec{H})$, i.e., that the magnetic flux density is a function of the applied magnetic field strength and is in the same direction. With the differential permeability defined as

$$\mu_d(H) \equiv \frac{\partial B}{\partial H}, \tag{3}$$

Equation (1) can be written as

$$\nabla \times \vec{E} = -\mu_d(H)\frac{\partial \vec{H}}{\partial t} \tag{4}$$

In general

$$B(H) = \mu_o[H + M(H)] \tag{5}$$

and

$$\mu_d(H) = \mu_o \; [1 + \frac{dM(H)}{dH}] \tag{6}$$

where

$$\mu_o = 4\pi \times 10^{-7} \text{ henry/meter} \tag{7}$$

is the permeability of free space and M(H) is the magnetization of the medium. As H is increased from zero, $\mu_d(H)$ for ferromagnetic materials usually starts from an initial value μ_i (which is the initial slope of the magnetization curve), increases to some maximum value, and then decreases to μ_o as the material undergoes saturation. For nonmagnetic materials M(H) is not present so that

$$B(H) = \mu_o H \tag{8}$$

and the differential permeability

$$\mu_d = \mu_o \tag{9}$$

is constant.

Now consider the unmagnetized, homogeneous, isotropic, conducting medium of semi-infinite extent ($x \geq o$) shown in Figure 1. At time t = o, a step increase in magnetic field strength in the $\hat{y}$ direction occurs at the surface x = o, e.g., as the result of an imposed surface current in the $\hat{z}$ direction. Since it is presumed that $H = H_y(x,t)\hat{y}$, the propagation of the electromagnetic fields following the onset of the step increase in $H_y(x,t)$ is governed by the equations

$$\frac{\partial E_z(x,\ t)}{\partial x} = \mu_d(H) \frac{\partial H_y(x,\ t)}{\partial t} \tag{10}$$

and

$$\frac{\partial H_y(x,\ t)}{\partial x} = \sigma E_z(x,\ t) \tag{11}$$

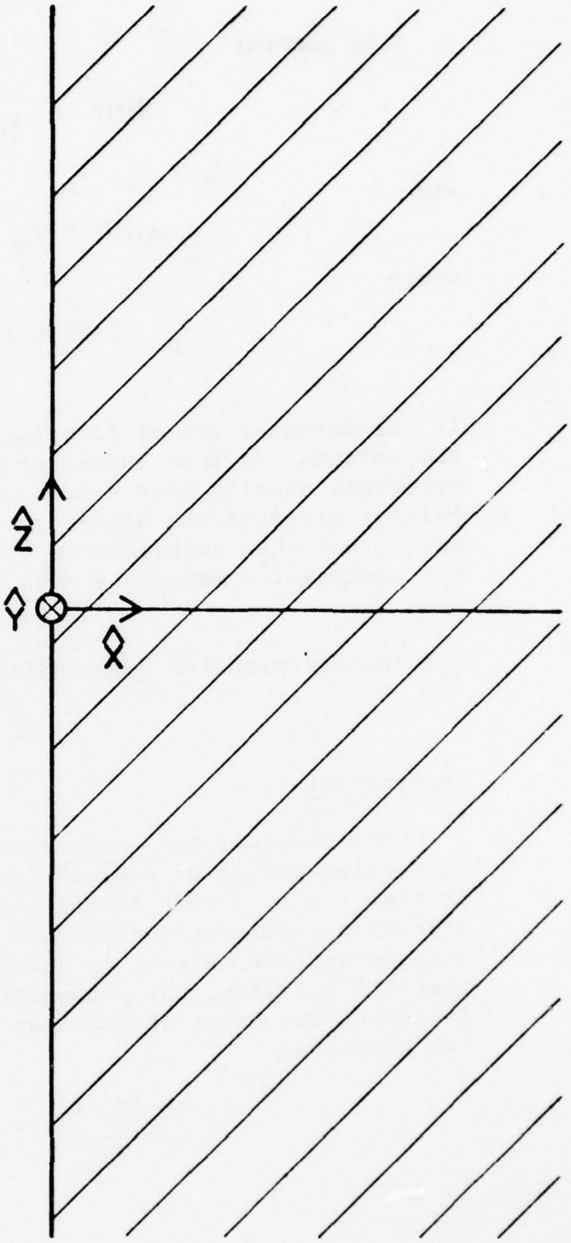


Figure 1. Semi-infinite medium subjected to a step increase in magnetic field strength at its surface. The magnetic field strength is in the $\hat{y}$ direction (into the page) and the associated electric field is in the $\hat{z}$ direction. Following the step change, the fields will propagate in the $\hat{x}$ direction into the medium.

which are obtained from (4) and (2) respectively. The partial differential equation governing the propagation of $H_y(x,t)$ is obtained by eliminating $E_z(x,t)$ from Equations (10) and (11).

The present problem requires the solution of the nonlinear partial differential equation

$$\frac{\partial^2 H_y(x, t)}{\partial x^2} = \sigma\mu_d(H) \frac{\partial H_y(x, t)}{\partial t} \tag{12}$$

subject to the initial condition

$$H_y(x, 0) = 0 \tag{13}$$

and the boundary conditions

$$H_y(0, t) = H_o \tag{14}$$

and

$$H_y(\infty, t) = 0. \tag{15}$$

The electric field can be determined from the solution to the above problem by noting from (11) that

$$E_z(x, t) = \frac{1}{\sigma} \frac{\partial H_y(x, t)}{\partial x} \tag{16}$$

REDUCTION IN THE NUMBER OF INDEPENDENT VARIABLES

Using an approach similar to that used for nonlinear diffusion phenomena analysis, the number of independent variables appearing in the problem can be reduced by the transformation

$$H_y(x, t) = H(\zeta) \tag{17}$$

where

$$\zeta = \frac{\sqrt{\sigma\mu_i}}{2} \frac{x}{\sqrt{t}} \tag{18}$$

The new variables are introduced by evaluating the partial derivatives in terms of the new variables and substituting these quantities into the original partial differential equation. On calculating the partial derivatives it is found that

$$\frac{\partial H_y(x,\, t)}{\partial x} = \frac{\partial \zeta}{\partial x}\,\frac{dH(\zeta)}{d\zeta} = \frac{\sqrt{\sigma\mu_i}}{2\sqrt{t}}\,\frac{dH(\zeta)}{d\zeta}\,, \tag{19}$$

$$\frac{\partial^2 H_y(x,\, t)}{\partial x^2} = \frac{\partial^2 \zeta}{\partial x^2}\,\frac{dH(\zeta)}{d\zeta} + \left(\frac{\partial \zeta}{\partial x}\right)^2 \frac{d^2 H(\zeta)}{d\zeta^2} \tag{20}$$

$$= \frac{\sigma\mu_i}{4t}\,\frac{d\,H(\zeta)}{d\zeta^2}\,,$$

and

$$\frac{\partial H_y(x,\, t)}{\partial t} = \frac{\partial \zeta}{\partial t}\,\frac{dH(\zeta)}{d\zeta} = -\,\frac{1}{2}\,\frac{\sqrt{\mu\sigma_i}}{2}\,\frac{x}{t^{3/2}}\,\frac{dH(\zeta)}{d\zeta}$$

$$= -\,\frac{1}{2}\,\frac{\zeta}{t}\,\frac{dH(\zeta)}{d\zeta}\,. \tag{21}$$

Upon substituting (20) and (21) into (12) and cancelling a factor of $\sigma\mu_i/4t$, it is found that the magnetic field strength satisfies the equation

$$\frac{d^2H(\zeta)}{d\zeta^2} = -2\zeta\,\frac{\mu_d(H)}{\mu_i}\,\frac{dH(\zeta)}{d\zeta}\,, \tag{22}$$

in which the variables x and t no longer appear explicitly. Noting that $x = o$ (and $t = \infty$) implies that $\zeta = o$, it follows from (14) that

$$H(o) = H_o \tag{23}$$

and noting that both $x = \infty$ and $t = o$ imply that $\zeta = \infty$, it follows from (13) and (15) that

$$H(\infty) = 0. \tag{24}$$

From (19) and (16) it is follows that the electric field is related to $H(\zeta)$ by

$$E_z(x, t) = \frac{1}{2}\sqrt{\frac{\mu_i}{\sigma}}\,\frac{1}{\sqrt{t}}\,\frac{dH(\zeta)}{d\zeta} \tag{25}$$

in which a factor of $\sqrt{t}$ appears. Alternatively, rearranging (24) yields the relation

$$E_z(x, t) = \frac{1}{\sigma x}\,\zeta\,\frac{dH(\zeta)}{d\zeta} \tag{26}$$

in which a factor of x appears. While the magnetic field strength is a function of ζ only, it is evident that the electric field is not a function of ζ alone.

For computational purposes, it is convenient to normalize the problem in terms of H_o. Define the normalized magnetic field function

$$F(\zeta) \equiv \frac{H(\zeta)}{H_o} = \frac{H_y(x, t)}{H_o} \tag{27}$$

which is simply the fraction of the applied value that the magnetic field strength has reached. In addition, it is convenient to introduce a nondimensional permeability function

$$P(F) \equiv \frac{\mu_d(H)}{\mu_i} = \frac{\mu_d(H_oF)}{\mu_i} \tag{28}$$

which is the differential permeability relative to μ_i expressed in terms of F, i.e., H is simply replaced by H_oF in the function $\mu_d(H)$.

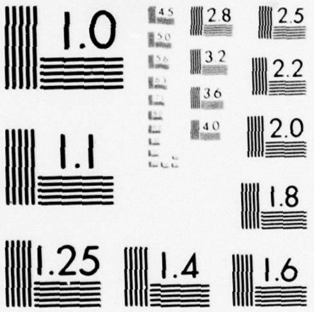
Thus, the problem given by (12) - (15) has been reduced to solving the nonlinear second order ordinary differential equation

$$\frac{d^2F}{d\zeta^2} = -2\zeta P(F)\,\frac{dF}{d\zeta} \tag{29}$$

AD-A064 433   ARMY RESEARCH OFFICE RESEARCH TRIANGLE PARK N C   F/G 12/1
TRANSACTIONS OF THE CONFERENCE OF ARMY MATHEMATICIANS(24TH) HEL--ETC(U)
JAN 79

UNCLASSIFIED   ARO-79-1   NL

3 OF 6
AD
A 064433



MICROCOPY RESOLUTION TEST CHART

NATIONAL BUREAU OF

subject to the conditions

$$F(0) = 1 \tag{30}$$

and

$$F(\infty) = 0 \tag{31}$$

The electric field can be evaluated from either of the relations

$$\frac{dF(\zeta)}{d\zeta} = \frac{2}{H_o}\sqrt{\frac{\sigma}{\mu_i}}\sqrt{t}\; E_z(x, t) \tag{32}$$

or

$$\zeta\frac{dF(\zeta)}{d\zeta} = \frac{\sigma x}{H_o} E_z(x, t) \tag{33}$$

Relation (32) is useful for evaluating the electric field distribution at a certain time whereas (33) is useful for examining the electric field variation with time at a particular location within the medium.

It should be emphasized that the transformation of variables (17) and (18) is not simply a change of variables resulting in a partial differential equation in terms of the new variables; rather, the number of independent variables is reduced by one, i.e., the magnetic field strength a function of ζ only. Therefore, it is evident that for a given differential permeability and applied H_o, the magnetic field strength will have the same value for all combinations of x and t which are related by

$$\zeta = \frac{\sqrt{\sigma\mu_i}}{2}\frac{x}{\sqrt{t}} = c \tag{34}$$

where c is a constant. Consequently, if a certain value for the magnetic field strength H has reached a location x at some time t, then the same value of H will occur at some location $x' = ax$ at a time $t' = a^2t$.

This is not the case for the electric field since the relations for the electric field (32) and (33) are not functions of the quantity ζ alone, i.e., there is explicit dependence on x or t.

SOLUTION FOR THE CONSTANT PERMEABILITY CASE

In classical analyses of electromagnetic field penetration, it is assumed that the permeability is constant:

$$\mu_d(H) = \mu_i. \tag{35}$$

In this case, the dimensionless permeability function is simply

$$P(F) = \frac{\mu_d}{\mu_i} = 1, \tag{36}$$

and the problem reduces to the linear equation

$$\frac{d^2F}{d\zeta^2} = -2\zeta \frac{dF}{d\zeta} \tag{37}$$

subject to the conditions

$$F(0) = 1 \tag{38}$$

and

$$F(\infty) = 0. \tag{39}$$

This equation can be integrated in a straightforward manner to obtain the classic solution

$$F(\zeta) = 1 - \mathrm{erf}(\zeta) = \mathrm{erfc}(\zeta) \tag{40}$$

where the error function is defined as

$$\mathrm{erf}(\zeta) = \frac{2}{\sqrt{\pi}} \int_0^\zeta \exp(-\zeta_1^2)\, d\zeta_1, \tag{41}$$

and the complementatary error function is defined as

$$\begin{aligned} \operatorname{erfc}(\zeta) &= 1 - \operatorname{erf}(\zeta) \\ &= \frac{2}{\sqrt{\pi}} \int_{\zeta}^{\infty} \exp(-\zeta_1^2) \, d\zeta_1 . \end{aligned} \tag{42}$$

It follows immediately from (32) and (33) that the relations pertaining to the electric field are

$$\frac{dF(\zeta)}{d\zeta} = -\frac{2}{\sqrt{\pi}} \exp(-\zeta^2) \tag{43}$$

and

$$\zeta \frac{dF(\zeta)}{d\zeta} = -\frac{2}{\sqrt{\pi}} \zeta \exp(-\zeta^2) . \tag{44}$$

PERTURBATION DUE TO SMALL CHANGES IN PERMEABILITY

The representation of the permeability as being constant, as in the classic solution considered in the preceding section, can be used as a first approximation in the case of a variable permeability. Small changes in permeability with magnetic field strength can be expected to introduce small perturbations of the classic solution for a constant permeability. If the permeability exhibits only a small variation with H over the range $0 \leq H \leq H_o$ then to a good approximation, the permeability can be represented by

$$\mu_d(H) = \mu_i \, (1 + aH) \tag{45}$$

where a is a constant that is characteristic of the material under consideration.

For the differential permeability given by (45), the non-dimensional permeability functions is

$$P(F) = 1 + \varepsilon F \tag{46}$$

where

$$\varepsilon = aH_o , \tag{47}$$

which is presumed to be a small quantity in the present analysis. In this case the problem requires the solution of the equation

$$\frac{d^2F}{d\zeta^2} = -2\zeta\,(1 + \varepsilon F)\,\frac{dF}{d\zeta} \tag{48}$$

subject to the conditions

$$F(0) = 1 \tag{49}$$

$$F(\infty) = 0 \tag{50}$$

A perturbation expansion in terms of the small parameter ε can now be used to examine the effects of small variations in permeability on the solution. To apply the perturbation method, a solution of the form

$$F(\zeta) = \sum_{m=0}^{\infty} \varepsilon^m f_m = f_0(\zeta) + \varepsilon f_1(\zeta) + \ldots \tag{51}$$

is assumed where the unknown functions $f_m(\zeta)$ are to be determined. Substitution of (51) into (48) - (50) yields, to first order terms in ε, the equation

$$\frac{d^2f_0}{d\zeta^2} + \varepsilon\,\frac{d^2f_1}{d\zeta^2} + \ldots = -2\zeta\frac{df_0}{d\zeta} - \varepsilon 2\zeta\left(\frac{df_1}{d\zeta} + f_0\,\frac{df_0}{d\zeta}\right) + \ldots \tag{52}$$

subject to the conditions

$$f_0(0) + \varepsilon f_1\,(0) + \ldots = 1 \tag{53}$$

and

$$f_0(\infty) + \varepsilon f_1(\infty) + \ldots = 0. \tag{54}$$

Since the above relations are to hold for any small value of ε, the coefficients of the various powers of ε are set equal.

By equating the quantities not involving ε, it is found that f_o must satisfy the equation

$$\frac{d^2 f_0}{d\zeta^2} = -2\zeta \frac{df_c}{d\zeta} \tag{55}$$

subject to the conditions

$$f_0(0) = 1 \tag{56}$$

and

$$f_0(\infty) = 0 \tag{57}$$

which is observed to be the classic problem for constant permeability (37) - (39). Thus it follows immediately that the solution for f_o is

$$f_0(\zeta) = \mathrm{erfc}(\zeta). \tag{58}$$

From the quantities involving first order terms in E, it is found that f_1 must satisfy the equation

$$\frac{d^2 f_1}{d\zeta^2} = -2\zeta \frac{df_1}{d\zeta} - 2\zeta f_0 \frac{df_0}{d\zeta} \tag{59}$$

subject to the conditions

$$f_1(0) = 0 \tag{60}$$

and

$$f_1(\infty) = 0 \tag{61}$$

The solution to this problem is

$$f_1(\zeta) = -\frac{1}{\pi}\mathrm{erfc}(\zeta) + \frac{1}{\pi}\exp(-2\zeta^2) - \frac{\zeta}{\sqrt{\pi}}\exp(-\zeta^2)\,\mathrm{erfc}(\zeta) \tag{62}$$

To first order terms inε , $F(\zeta)$ is given by (51) with f_o given by (58) and f_1 given by (62).

Similarly, to first order terms

$$\frac{dF(\zeta)}{d\zeta} = \frac{df_o}{d\zeta} + \varepsilon \frac{df_1}{d\zeta} + \ldots \tag{63}$$

where

$$\frac{df_o(\zeta)}{d\zeta} = -\frac{2}{\sqrt{\pi}} \exp(-\zeta^2) \tag{64}$$

and

$$\frac{df_1(\zeta)}{d\zeta} = \frac{2}{\pi^{3/2}} \exp(-\zeta^2) - \frac{2}{\pi} \zeta \exp(-2\zeta^2) \tag{65}$$

$$+ \frac{2}{\sqrt{\pi}} \zeta^2 \exp(-\zeta^2) \operatorname{erfc}(\zeta) - \frac{1}{\sqrt{\pi}} \exp(-\zeta^2) \operatorname{erfc}(\zeta).$$

Likewise

$$\zeta \frac{dF(\zeta)}{d\zeta} = \zeta \frac{df_o}{d\zeta} + \varepsilon \, \zeta \frac{df_1}{d\zeta} + \ldots \tag{66}$$

where

$$\zeta \frac{df_o}{d\zeta} = -\frac{2}{\sqrt{\pi}} \zeta \exp(-\zeta^2) \tag{67}$$

and

$$\zeta \frac{df_1}{d\zeta} = \frac{2}{\pi^{3/2}} \zeta \exp(-\zeta^2) - \frac{2}{\pi} \zeta^2 \exp(-2\zeta^2) \tag{68}$$

$$+ \frac{2}{\sqrt{\pi}} \zeta^3 \exp(-\zeta^2) \operatorname{erfc}(\zeta) - \frac{1}{\sqrt{\pi}} \zeta \exp(-\zeta^2) \operatorname{erfc}(\zeta).$$

The quantities appearing in the above perturbation analysis are tabulated in Table 1. The perturbation functions f_1, $df_1/d\zeta$, and $\zeta\, df_1/d\zeta$ are also plotted in Figures 2, 3, and 4.

It is of interest to note that the variation of P(F) given by (46) results in a perturbation of $dF(0)/d\zeta$ which to first order terms in ε is given by

$$\frac{dF(0)}{d\zeta} = \frac{-2}{\sqrt{\pi}}\left[1 + \varepsilon\left(\frac{1}{2} - \frac{1}{\pi}\right) + \ldots\right] \tag{69}$$

FORMAL SOLUTION FOR AN EXPONENTIALLY DECREASING PERMEABILITY

In most ferromagnetic materials which have not completely saturated, the variation of B with H is primarily due to the variation of M(H). Although the exact magnetization curve can be somewhat complicated over the range of H from zero to that required for saturation, for cases of H below that for complete saturation the major variation of B(H) can be approximated reasonably well by the simple representation

$$B(H) = B_s[1 - \exp(-H/H_m)] \tag{70}$$

where B_s and H_m are parameters (to be determined from empirical data) which govern the magnitude and shape of the magnetization curve for the medium over the range of interest. With this representation of the magnetization curve, the differential permeability is given by

$$\mu_d(H) = \mu_i \exp(-H/H_m) \tag{71}$$

where

$$\mu_i = B_s/H_m \tag{72}$$

is the initial slope of the B-H curve given by (70).

The variation of B with H for this approximation is shown in Figure 5 and the corresponding variation of μ_d with H is shown in Figure 6. As the material saturates at large values of H, an actual

ζ	f_o	f_1	$df_o/d\zeta$	$df_1/d\zeta$	$\zeta df_o/d\zeta$	$\zeta df_1/d\zeta$
0.0	1.00000	-0.00000	-1.12838	-0.20502	-0.00000	-0.00000
0.1	0.88754	-0.02008	-1.11715	-0.19264	-0.11172	-0.01926
0.2	0.77730	-0.03785	-1.08413	-0.16008	-0.21683	-0.03202
0.3	0.67137	-0.05168	-1.03126	-0.11513	-0.30938	-0.03454
0.4	0.57161	-0.06073	-0.96154	-0.06572	-0.38462	-0.02629
0.5	0.47950	-0.06491	-0.87878	-0.01868	-0.43939	-0.00934
0.6	0.39614	-0.06472	-0.78724	0.02100	-0.47235	0.01260
0.7	0.32220	-0.06105	-0.69127	0.05056	-0.48389	0.03539
0.8	0.25790	-0.05497	-0.59499	0.06927	-0.47599	0.05542
0.9	0.20309	-0.04753	-0.50197	0.07800	-0.45177	0.07020
1.0	0.15730	-0.03964	-0.41511	0.07862	-0.41511	0.07862
1.1	0.11979	-0.03200	-0.33648	0.07345	0.37013	0.08080
1.2	0.08969	-0.02507	-0.26734	0.06475	-0.32081	0.07770
1.3	0.06599	-0.01910	-0.20821	0.5445	-0.27067	0.07078
1.4	0.04771	-0.01418	-0.15894	0.04398	-0.22252	0.06157
1.5	0.03389	-0.01028	-0.11893	0.03430	-0.17840	0.05145
1.6	0.02365	-0.00728	-0.08723	0.02593	-0.13957	0.04149
1.7	0.01621	-0.00504	-0.06271	0.01905	-0.10661	0.03238
1.8	0.01091	-0.00342	-0.04419	0.01363	-0.07955	0.02453
1.9	0.00721	-0.00227	-0.03052	0.00952	-0.05800	0.01808
2.0	0.00468	-0.00148	-0.02067	0.00648	-0.04133	0.01298
2.1	0.00298	-0.00094	-0.01372	0.00433	-0.02880	0.00909
2.2	0.00186	-0.00059	-0.00892	0.00282	-0.01963	0.00621
2.3	0.00114	-0.00036	-0.00569	0.00180	-0.01308	0.00415
2.4	0.00069	-0.00022	-0.00356	0.00113	-0.00853	0.00271
2.5	0.00041	-0.00013	-0.00218	0.00069	-0.00545	0.00173
2.6	0.00024	-0.00008	-0.00131	0.00042	-0.00340	0.00108
2.7	0.00013	-0.00004	-0.00077	0.00024	-0.00208	0.00066
2.8	0.00008	-0.00002	-0.00044	0.00014	-0.00124	0.00040
2.9	0.00004	-0.00001	-0.00025	0.00008	-0.00073	0.00023
3.0	0.00002	-0.00001	-0.00014	0.00004	-0.00042	0.00013

Table 1. Calculated Values of Functions f_o, f_1, $df_o/d\zeta$, $\zeta\, df_o/d\zeta$, and $\zeta df_1/d\zeta$.

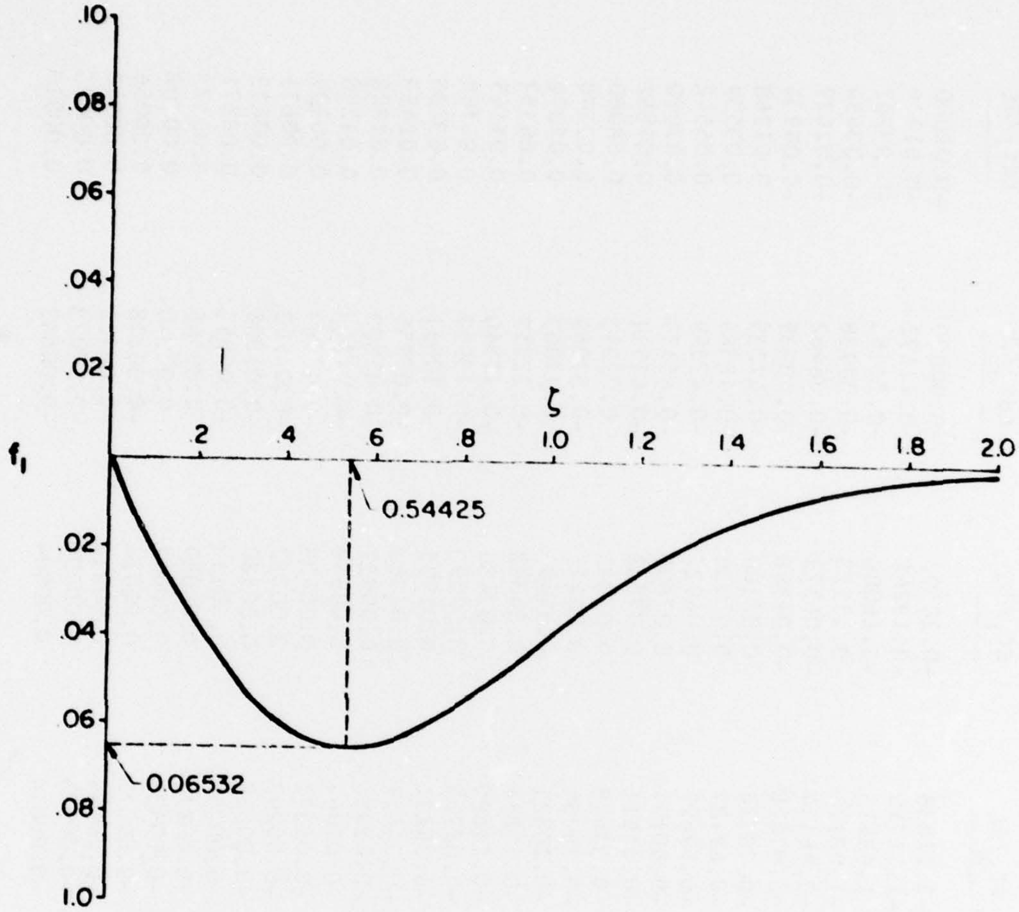


Figure 2. First order perturbation function f_1.

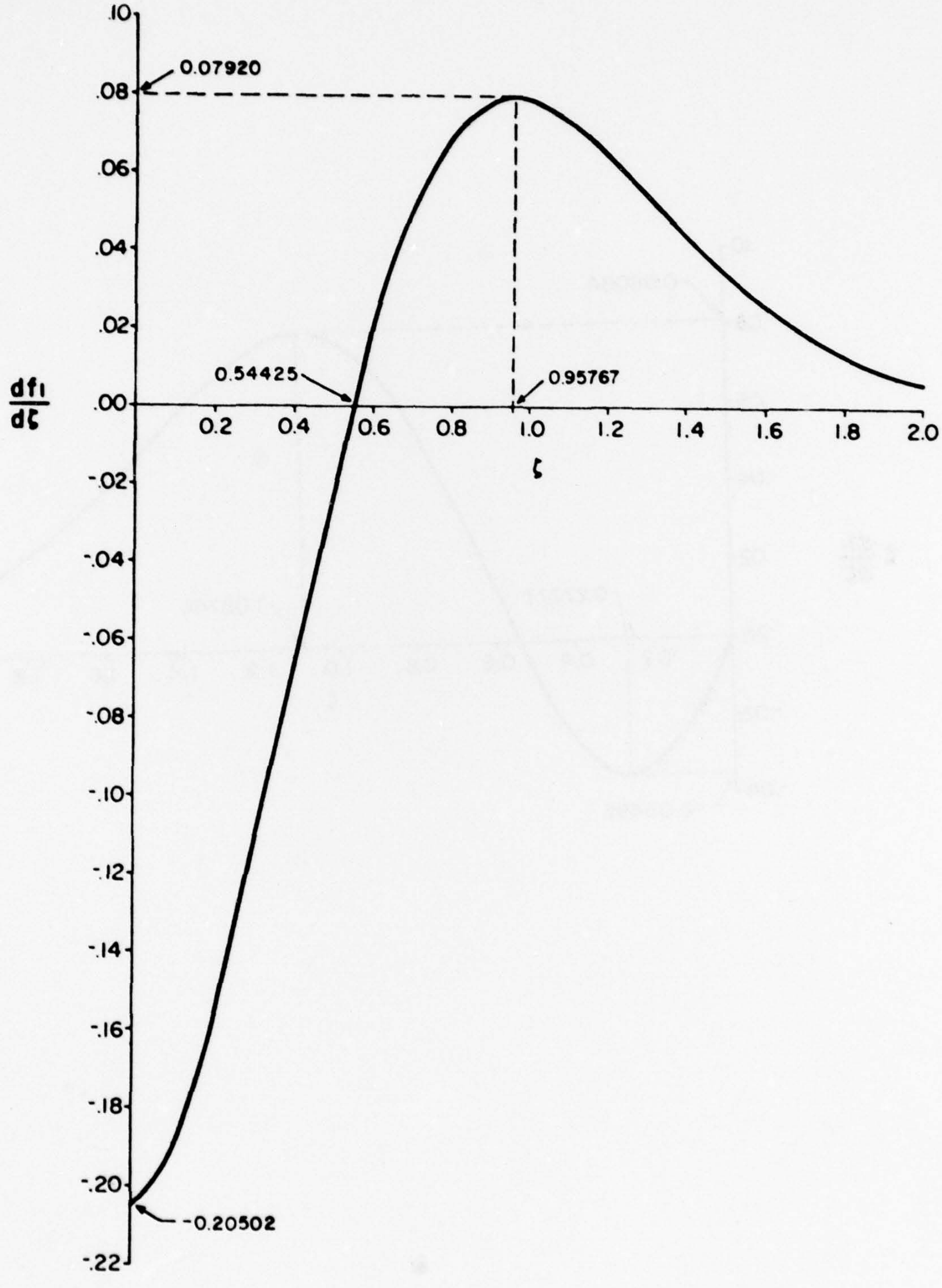


Figure 3. First order perturbation function $\frac{df_1}{d\zeta}$.

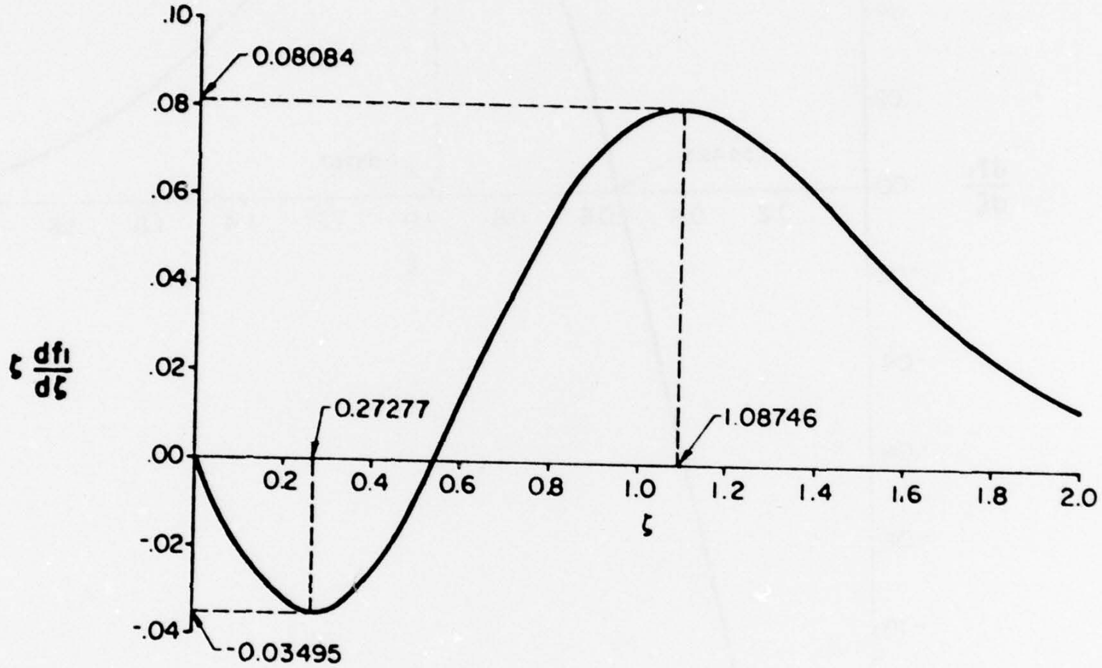


Figure 4. First order perturbation function $\zeta\frac{df_1}{d\zeta}$.

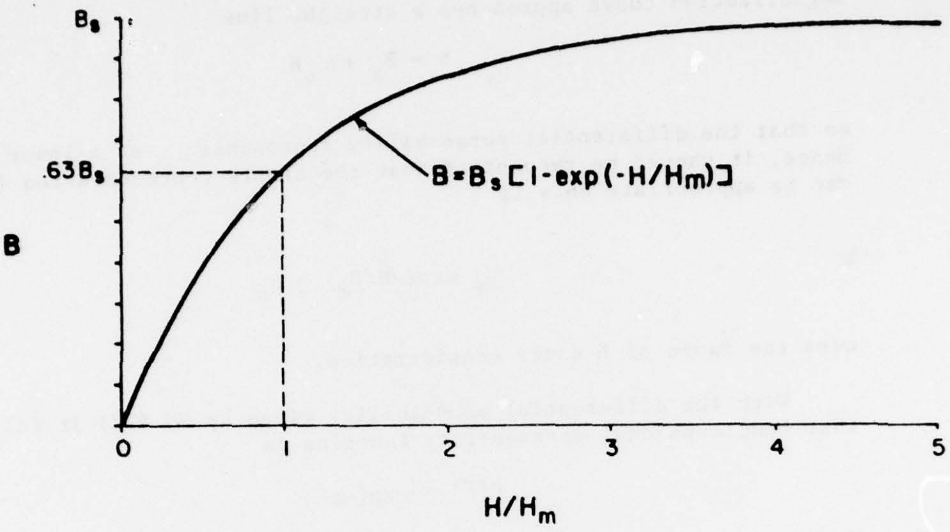


Figure 5. Idealized magnetization curve.

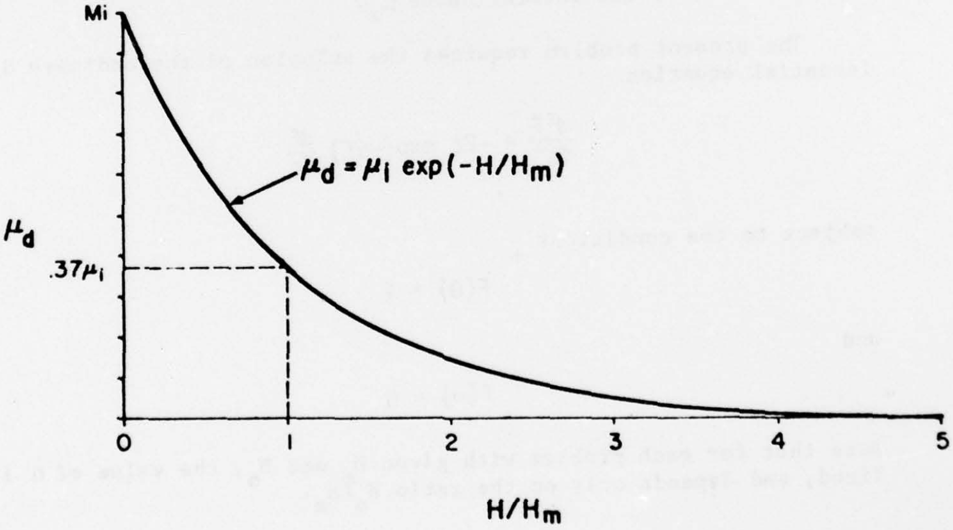


Figure 6. Idealized differential permeability.

magnetization curve approaches a straight line

$$B = B_s + \mu_o H$$

so that the differential permeability approaches μ_o as a lower limit. Hence, it should be recognized that the simple representation (71) can be appropriate only if

$$\mu_i \exp(-H/H_m) \geq \mu_o$$

over the range of H under consideration.

With the differential permeability given by Eq (72) it follows that nondimensional permeability function is

$$P(F) = \exp(-\alpha F) \tag{73}$$

where

$$\alpha = H_o/H_m \tag{74}$$

The saturation index α is a parameter which indicates the degree of saturation (or the extent of tne decrease in permeability for $H = H_o$) For example, $\alpha = 1$ indicates that the permeability at $H = H_o = H_m$ would be 0.37 of the initial value μ_i.

The present problem requires the solution of the ordinary differential equation

$$\frac{d^2F}{d\zeta^2} = -2\zeta \exp(-\alpha F) \frac{dF}{d\zeta} \tag{75}$$

subject to the conditions

$$F(0) = 1 \tag{76}$$

and

$$F(\infty) = 0 \tag{77}$$

Note that for each problem with given H_m and H_o, the value of α is fixed, and depends only on the ratio H_o/H_m.

Most of the standard linear techniques are of relatively little use in achieving an exact solution in nonlinear cases ($\alpha \neq 0$); however, in the course of an investigation of a related problem in nonlinear diffusion phenomena, Fujita [1] found a formal solution which can be directly adapted to the problem presently under investigation. The problem is satisfied with F and ζ being given by the parametric representations

$$F = \frac{2}{\alpha}\int_0^{\xi}[\xi_1^2 - \beta\ell n(\xi_1^2)]^{-1/2}\, d\xi_1 \qquad (78)$$

$$\zeta = \frac{1}{\sqrt{2\beta}}\{[\xi^2 - \beta\ell n(\xi^2)]^{1/2} - \xi\} \times \exp\{\int_0^{\xi}[\xi_1^2 - \beta\ell n(\xi_1)^2]^{-1/2}\, d\xi_1\}, \qquad (79)$$

where the derived quantity β is related to the given quantity α by

$$\alpha = 2\int_0^1[\xi_1^2 - \beta\ell n\ (\xi_1^2)]^{-1/2}\, d\xi_1 \qquad (80)$$

The parameter ξ ($0 \leq \xi \leq 1$) relates a value of F to the corresponding value of ζ. (Note that $\xi = 0$ implies $\zeta = \infty$ and $F = 0$, while $\xi = 1$ implies $\zeta = 0$ and $F = 1$.) To evaluate F versus ζ, a value of ξ is selected and the value of F and the corresponding value of ζ are evaluated.

The parametric representation of $dF/d\zeta$ can be evaluated from the above solution for F through the relation:

$$\frac{dF}{d\zeta} = \left(\frac{dF}{d\xi}\right)\Big/\left(\frac{d\zeta}{d\xi}\right)$$

It follows that

$$\frac{dF}{d\zeta} = -\frac{2\sqrt{2}}{\alpha\sqrt{\beta}}\ \xi\ \exp\{-\int_0^{\xi}[\xi_1^2 - \beta\ell n(\xi_1^2)]^{-1/2}\, d\xi_1\} \qquad (81)$$

The parametric representation of $\zeta\, dF/d\zeta$ is

$$\zeta\frac{dF}{d\zeta} = -\frac{2}{\alpha\beta}\ \xi\{[\xi^2 - \beta\ell n(\xi^2)]^{1/2} - \xi\} \tag{82}$$

The integrals appearing in the formal solution do not appear to be evaluated in closed form, thus numerical integration seems to be required; however, the above representation offers the advantage of allowing the computational effort to be directed to the very accurate calculation of the solution at only a few points, or one point for that matter, since the knowledge of intermediate points is not required. Of special interest is $dF/d\zeta$ evaluated at $\zeta = 0$ where

$$\frac{dF(0)}{d\zeta} = -\frac{2\sqrt{2}}{\alpha\sqrt{\beta}}\exp(-\alpha/2)\ . \tag{83}$$

In addition to its relation to the electric field at $x = 0$, the calculated value of $dF(0)/d\zeta$ can be used to transform the problem from a two-point problem (with a condition at $x = \infty$ as well as one at $x = 0$) to an initial value problem (with both boundary conditions specified at $x = 0$). Such a transformation is often useful from a computational standpoint.

ASSOCIATED INITIAL VALUE PROBLEM

In some situations the formal solution to the problem may not be the most convenient form to generate the desired information. In the study of pulse penetration phenomena, it is frequently of interest to know the time that it takes for the electromagnetic fields at a certain location in the medium to rise to a certain fraction of their maximum value or to know the position to which a specified field level has penetrated at a prescribed time. In other words, the primary objective may be to determine the variation with applied field level of the value of ζ at which the solution F has a prescribed value. Even in those cases in which the solution can be expressed in terms of elementary functions, the formal solution to the problem may be unwieldly, and, in those cases involving numerical calculations, repeated iterations may be cumbersome and undesirable.

If F(ζ) is a known function of ζ and F as well as all its derivatives are continuous and well-behaved in the neighborhood of the point ζ = 0, then F(ζ) can be expanded in a Maclaurin power series about ζ = 0:

$$F(\zeta) = \sum_{m=0}^{\infty} \frac{d^mF(0)}{d\zeta^m} \frac{\zeta^m}{m!} = F(0) + \frac{dF(0)}{d\zeta}\zeta + \frac{d^2F(0)}{d\zeta^2}\frac{\zeta^2}{2!} + \ldots \tag{84}$$

where F and its derivatives are evaluated at ζ = 0. For example, the solution to the linear case (39) has the well-known series representation

$$F(\zeta) = 1 - \frac{2}{\sqrt{\pi}} \sum_{m=0}^{\infty} \frac{(-1)^m \zeta^{2m+1}}{(2m+1)m!} = 1- \frac{2}{\sqrt{\pi}}\left(\zeta - \frac{\zeta^3}{3} + \frac{\zeta^5}{5\cdot 2!} - \frac{\zeta^7}{7\cdot 3!} + \ldots\right) \tag{85}$$

If the Maclaurin series expansion converges rapidly, the sum of the first few terms gives a good approximation to F(ζ) for values of near ζ = 0.

A series expansion can also be used to develop a series representation of the solution F(ζ) from the second order differential equation -- if F(0) and dF(0)/dζ are given as boundary conditions. In such an initial value problem, the approach is straight-forward since the second and higher order derivatives can, in principle, be determined by successive differentiation of the second order ordinary differential equation.

Although F(0) = 1 is one of the boundary conditions in the problem presently under investigation, it is evident that a difficulty arises because the second condition F(∞) = 0 is specified at ζ = ∞. For this reason we are lead to consider the associated initial value problem requiring the solution of the equation

$$\frac{d^2F}{d\zeta^2} = -2\zeta P(F)\frac{dF}{d\zeta} \tag{86}$$

subject to

$$F(0) = 1 \tag{87}$$

and

$$\frac{dF(0)}{d\zeta} = \gamma , \tag{88}$$

where the value for γ is to be chosen such that F (∞) = 0. In general such a value for γ is not known for an arbitrary P(F); however, in some cases, the value for γ can be determined from other analytical considerations, e.g. Eq (83) can be used to compute γ for an exponentially decreasing permeability.

The second derivative at $\zeta = 0$ can be determined directly from the governing differential equation (86) to be

$$\frac{d^2F(0)}{d\zeta^2} = 0 \tag{89}$$

By differentiation of Eq (86), the third derivative is found to be

$$\frac{d^3F(\zeta)}{d\zeta^3} = -2 \{P(F) \frac{dF(\zeta)}{d\zeta} + \zeta\frac{d}{d\zeta} [P(F) \frac{dF(\zeta)}{d\zeta}]\}, \tag{90}$$

Since F(0) = 1 and $dF(0)/d\zeta = \gamma$ it follows that

$$\frac{d^3F(0)}{d\zeta^3} = -2\gamma P(1) \tag{91}$$

Differentiating (38) and evaluating at $\zeta = 0$ yields

$$\frac{d^4F}{d\zeta^4}(0) = -4\gamma^2 P'(1) \tag{92}$$

where P′(1) indicates dP/dF evaluated at F(0) = 1. Similarly

$$\frac{d^5F}{d\zeta^5}(0) = -6 [\gamma^3 P''(1) -2 P\gamma(1)], \tag{93}$$

$$\frac{d^6F}{d\zeta^6}(0) = -8 [\gamma^4 P'''(1) - 12\gamma^2 P(1) P'(1)], \tag{94}$$

and

$$\frac{d^7F(0)}{d\zeta^7} -10[\gamma^5P^{(IV)}(1) - 20\gamma^3(P'(1))^2 - 26\gamma^3P(1)P''(1) + 12\gamma P^3(1)] \qquad (95)$$

where $P^{(n)}(1)$ denotes $d^nP/d\zeta^n$ evaluated at $F(0) = 1$. The approach could in prinicple be extended indefinitely; however, it becomes rather tedious after the first few terms.

Substituting (87) - (95) into (84) yields the Maclaurin series expansion at $F(\zeta)$ to seventh order terms:

$$\begin{aligned} F(\zeta) = 1 &+ \gamma\zeta - \frac{1}{3}\gamma P(1)\,\zeta^3 - \frac{1}{6}\gamma^2P'(1)\zeta^4 \\ &- \frac{1}{20}[\gamma^3P''(1) - 2\gamma P^2(1)]\zeta^5 \\ &- \frac{1}{90}[\gamma^4P'''(1) - 12\gamma^2P(1)P'(1)]\zeta^6 \\ &- \frac{1}{504}[\gamma^5P^{(IV)}(1) - 20\gamma^3(P'(1))^2 - 26\gamma^3P(1)\,P''(1) + 12\gamma P^3(1)]\zeta^7 \\ &\dots \end{aligned} \qquad (96)$$

Although care must be exercised when using these formulas to verify that the series converges, such series representations can be useful for calculating the solution when the terms after the first few may be neglicted (for example, if ζ is much less than unity.)

As previously mentioned, it is frequently of interest to know the value of ζ for which F has a prescribed value. For this purpose a convenient manipulation is the reversion of series. Given a series for a dependent variable w in terms of an independent variable u:

$$w = a_1u + a_2u^2 + a_3u^3 + a_4u^4 + a_5u^5 + a_6u^6 + a_7u^7 + \dots$$

a reversion of series can be used to write a series for u in terms of w:

$$u = c_1 w + c_2 w^2 + c_3 w^3 + c_4 w^4 + c_5 w^5 + c_6 w^6 + c_7 w^7 + \cdots$$

where

$$c_1 = \frac{1}{a_1},$$

$$c_2 = -\frac{a_2}{a_1^3},$$

$$c_3 = \frac{1}{a_1^5}(2a_2^2 - a_1 a_3),$$

$$c_4 = \frac{1}{a_1^7}(5a_1a_2a_3 - a_1^2a_4 - 5a_2^3),$$

$$c_5 = \frac{1}{a_1^9}(6a_1^2a_2a_4 + 3a_1^2a_3^2 + 14a_2^2 - a_1^3a_5 - 21a_1a_2^2a_3),$$

$$c_6 = \frac{1}{a_1^{11}}(7a_1^3a_2a_5 + 7a_1^3a_3a_4 + 84a_1a_2^3a_3 - a_1^4a_6 - 28a_1^2a_2a_3^2 -$$
$$42a_2^5 - 28a_1^2a_2^2a_4),$$

$$c_7 = \frac{1}{a_1^{13}}(8a_1^4a_2a_6 + 8a_1^4a_3a_5 + 4a_1^4a_4^2 + 120a_1^2a_2^3a_4 + 180a_1^2a_2^2a_3^2 + 132a_2^6 -$$
$$a_1^5a_7 - 36a_1^3a_2^2a_5 - 72a_1^3a_2a_3a_4 - 12a_1^3a_3^3 - 330a_1a_2^4a_3),$$

$$\cdots$$

The series expansion (96) can be rewritten

$$(1 - F) = -\gamma\zeta + \frac{1}{3}\gamma P(1)\zeta^3 + \frac{1}{6}\gamma^2 P'(1)\zeta^4$$

$$+ \frac{1}{20}[P''(1)\gamma^3 - 2P^2(1)\gamma]\zeta^5$$

$$+ \frac{1}{90}[P'''(1)\gamma^4 - 12P(1)P'(1)\gamma^2]\zeta^6$$

$$+ \frac{1}{504}[P^{(IV)}(1)\gamma^5 - 20(P'(1))^2\gamma^3 - 26P(1)P''(1)\gamma^3 + 12P^3(1)\gamma]\zeta^7$$

$$+ \ldots$$

A reversion of series yields

$$\zeta = -\frac{1}{\gamma}\{(1 - F) + \frac{P(1)}{3\gamma^2}(1 - F)^3 - \frac{1}{6}\frac{P'(1)}{\gamma^2}(1 - F)^4 \qquad (97)$$

$$+ [\frac{7}{30}\frac{P^2(1)}{\gamma^4} + \frac{1}{20}\frac{P''(1)}{\gamma^2}](1 - F)^5$$

$$- [\frac{23}{90}\frac{P(1)P'(1)}{\gamma^4} + \frac{1}{90}\frac{P'''(1)}{\gamma^2}](1 - F)^6$$

$$+ [\frac{1}{504}\frac{P^{(IV)}(1)}{\gamma^2} + \frac{103}{1260}\frac{P(1)P''(1)}{\gamma^4} + \frac{1}{14}\frac{(P'(1))^2}{\gamma^4} + \frac{127}{630}\frac{P^3(1)}{\gamma^6}](1 - F)^7$$

$$+ \ldots\}$$

which can be used to examine the variation of the value of ζ at which $F(\zeta)$ has a specified value for [for small values of (1-F)].

In the case of an exponentially decreasing permeability

$$P(1) = \exp(-\alpha)$$

$$P'(1) = -\alpha\exp(-\alpha)$$

$$P''(1) = (-\alpha)^2\exp(-\alpha)$$

$$\vdots$$

$$P^{(n)}(1) = (-\alpha)^n \exp(-\alpha)$$

Thus, the series expansion of F for an exponentially decreasing permeability is given by

$$F(\zeta) = 1 + \gamma\zeta - \frac{1}{3}\gamma\exp(-\alpha)\zeta^3 + \frac{1}{6}\gamma^2\alpha\exp(-\alpha)\zeta^4 \tag{98}$$

$$- \frac{1}{20}[\gamma\alpha^2\exp(-\alpha) - 2\gamma\exp(-2\alpha)]\zeta^5$$

$$+ \frac{1}{90}[\gamma^4\alpha^3\exp(-\alpha) - 12\gamma^2\alpha\exp(-2\alpha)]\zeta^6$$

$$- \frac{1}{504}[\gamma^5\alpha^4\exp(-\alpha) - 46\gamma^3\alpha^2\exp(-2\alpha) + 12\gamma\exp(-3\alpha)]\zeta^7 + \ldots$$

where γ is determined from Eq (83).

In this case, a reversion of series yields

$$\zeta = -\frac{1}{\gamma}\{(1 - F) + \frac{\exp(-\alpha)}{3\gamma^2}(1 - F)^3 + \frac{\alpha\exp(-\alpha)}{6\gamma^2}(1 - F)^4 \tag{99}$$

$$+ [\frac{7\alpha\exp(-2\alpha)}{30\gamma^4} + \frac{\alpha^2\exp(-\alpha)}{20\gamma^2}](1 - F)^5$$

$$+ [\frac{23\alpha\exp(-2\alpha)}{90\gamma^4} + \frac{\alpha^3\exp(-\alpha)}{90\gamma^2}](1 - F)^6$$

$$+ [\frac{\alpha^4\exp(-\alpha)}{504\gamma^2} + \frac{193\alpha^2\exp(-2\alpha)}{1260\gamma^4} + \frac{127\exp(-3\alpha)}{630\gamma^6}](1 - F)^7$$

$$+ \ldots\}$$

which is suitable for calculating the value of ζ at which the solution $F(\zeta)$ has a specified value [for small values of $(1-F)$, which implies small ζ].

NUMERICAL RESULTS FOR AN EXPONENTIALLY DECREASING PERMEABILITY

The parametric solutions for the case of an exponentially decreasing permeability were given in Eqs (78) - (82) where the parameter ξ relates a value for F, $dF/d\zeta$, and $\zeta dF/d\zeta$ to a corresponding value of ζ. Since the integral

$$\int_0^{\xi} [\xi_1^2 - \beta \ln(\xi_1^2)]^{-1/2} \, d\xi_1$$

does not appear to have been evaluated in closed form, numerical integration is required. The relations (78) - (82) can be used for numerical calculation of the solution; however, the approach is somewhat inconvenient because the calculation at each point requires the numerical integration of the aforementioned integral. Moreover, it is difficult to determine from Eq (80) the value of the derived parameter β which corresponds to the given parameter α. On the other hand, the value of α for a specified value of β can be computed in a straight-forward manner by numerical integration of (80), and the value of γ can be calculated from (83). This value of γ can then be used in the series expansion (96) or in a numerical algorithm such as the Runge-Kutta as an alternative way to generate the solution $F(\zeta)$.

Some numerical calculations were performed and the results are shown in Table 2. An examination of Table 2 reveals that rather large changes in β correspond to relatively small changes in α. This situation can make it difficult to determine the value of β which corresponds to a specified value of α. On the other hand, it is evident that $dF(0)/d\zeta = \gamma$ exhibits a relatively moderate variation with α. While α varies in the range $0 \leq \alpha < \infty$, γ varies in the range $-2/\sqrt{\pi} \leq \gamma < 0$. This moderate variation makes it possible to accurately represent γ as a function of α using the regression equation

$$\begin{aligned} \frac{dF(0)}{d\zeta} = \gamma = & - 1.1284 + 2.0490 \times 10^{-1}\alpha - 3.4426 \times 10^{-2}\alpha^2 \\ & + 4.4139 \times 10^{-3}\alpha^3 - 4.1562 \times 10^{-4}\alpha^4 \\ & + 2.7025 \times 10^{-5}\alpha^5 - 1.0736 \times 10^{-6}\alpha^6 \\ & + 0.9395 \times 10^{-7}\alpha^7 - \ldots, \end{aligned} \qquad (100)$$

which provides a convenient means of evaluating γ for $0 \leq \alpha < 10$. A plot of γ versus α is shown in Figure 7. For small values of α, the first two terms on the right hand side of (100) are equivalent to the right hand side of (69) if ϵ is replaced by $-\alpha$.

Table 2

Numerical Calculations of α and γ.*

β	α	$\gamma = \dfrac{dF(0)}{d\zeta}$
∞	0	$-2/\sqrt{\pi}$
5×10^4	0.0112	-1.1260
1×10^4	0.0249	-1.1233
5×10^3	0.0351	-1.1212
1×10^3	0.0773	-1.1127
5×10^2	0.1083	-1.1066
1×10^2	0.2326	-1.0825
5×10^1	0.3197	-1.0662
1×10^1	0.6424	-1.0098
5×10^0	0.8475	-0.9769
1×10^0	1.5058	-0.8847
5×10^{-1}	1.8681	-0.8414
1×10^{-1}	2.8683	-0.7432
5×10^{-2}	3.3552	-0.7043
1×10^{-2}	4.5797	-0.6255
5×10^{-3}	5.1378	-0.5965
1×10^{-3}	6.4838	-0.5392
5×10^{-4}	7.0801	-0.5183
1×10^{-4}	8.4931	-0.4767
5×10^{-5}	9.1115	-0.4612
1×10^{-5}	10.5651	-0.4300

*The calculations were performed on a CDC 6600 computer using simple trapezoidal numerical integration.

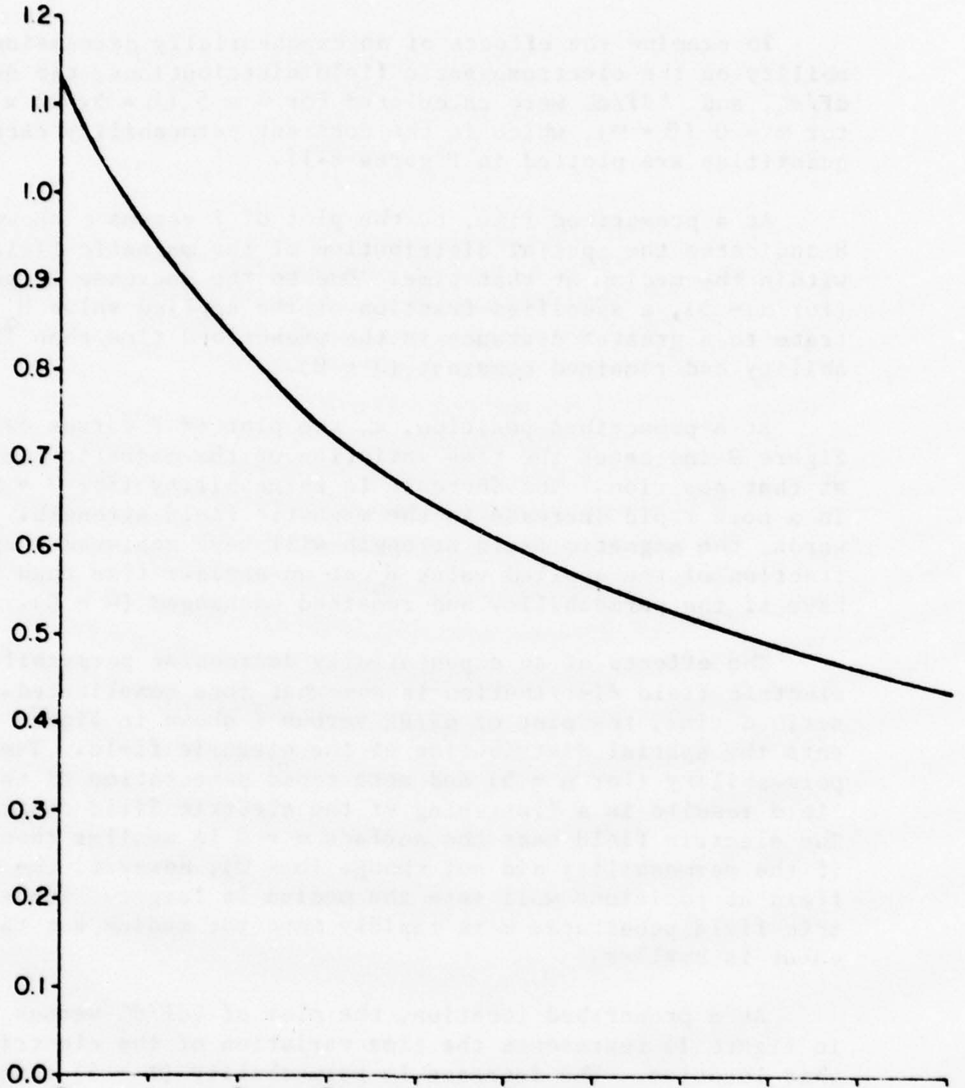


Figure 7. Initial slope (γ) of formal solution vs Saturation Index ( ).

To examine the effects of an exponentially decreasing permeability on the electromagnetic field distributions, the quantities F, $dF/d\zeta$, and $\zeta dF/d\zeta$ were calculated for $\alpha = 5$ ($\beta = 5.922 \times 10^{-3}$) and for $\alpha = 0$ ($\beta = \infty$), which is the constant permeability case. These quantities are plotted in Figures 8-11.

At a prescribed time, t, the plot of F versus ζ shown in Figure 8 indicates the spatial distribution of the magnetic field strength within the medium at that time. Due to the decrease in permeability (for $\alpha = 5$), a specified fraction of the applied value H_o will penetrate to a greater distance in the prescribed time than if the permeability had remained constant ($\alpha = 0$).

At a prescribed position, x, the plot of F versus $1/\zeta^2$ shown in Figure 9 indicates the time variation of the magnetic field strength at that position. The decrease in permeability (for $\alpha = 5$) results in a more rapid increase in the magnetic field strength. In other words, the magnetic field strength will have achieved a specified fraction of the applied value H_o at an earlier time than it would have if the permeability had remained unchanged ($\alpha = 0$).

The effects of an exponentially decreasing permeability on the electric field distribution is somewhat more complicated. For a prescribed time, the plot of $dF/d\zeta$ versus ζ shown in Figure 10 represents the spatial distribution of the electric field. The decrease in permeability (for $\alpha = 5$) and more rapid penetration of the magnetic field results in a flattening of the electric field distribution. The electric field near the surface $x = 0$ is smaller than it would be if the permeability did not change ($\alpha = 0$); however, the electric field at positions well into the medium is larger. Hence, the electric field penetrates more rapidly into the medium but the maximum value is smaller.

At a prescribed location, the plot of $\zeta dF/d\zeta$ versus $1/\zeta^2$ shown in Figure 11 represents the time variation of the electric field at that location. The decrease in permeability ($\alpha = 5$) results in the electric field increasing to its maximum value earlierand achieving a larger maximum value than if the permeability had not decreased ($\alpha = 0$);however, the electric field decreases more rapidly and ultimately is less than that for a constant permeability. The quantity $\zeta dF/d\zeta$ is asymptotic to zero as $1/\zeta^2 \to \infty$.

The magnetic field strength at any point within the medium is asymptotic to the value applied at the surface, however since the medium is presumed to be infinite in extent, the magnetic field

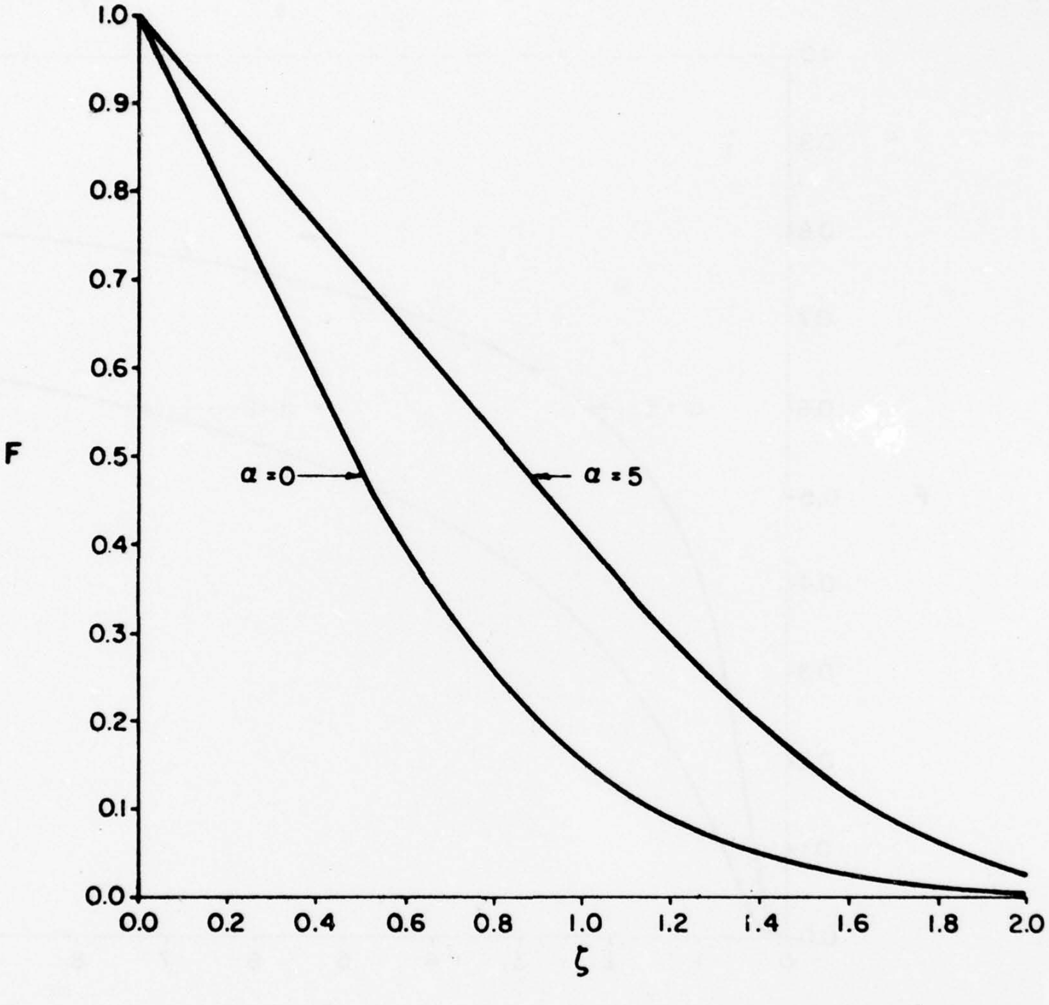


Figure 8. Normalized magnetic field function F versus ζ for α=0 and α=5. For a fixed time ζ is proportional to x and the above plot yields the normalized magnetic field distribution.

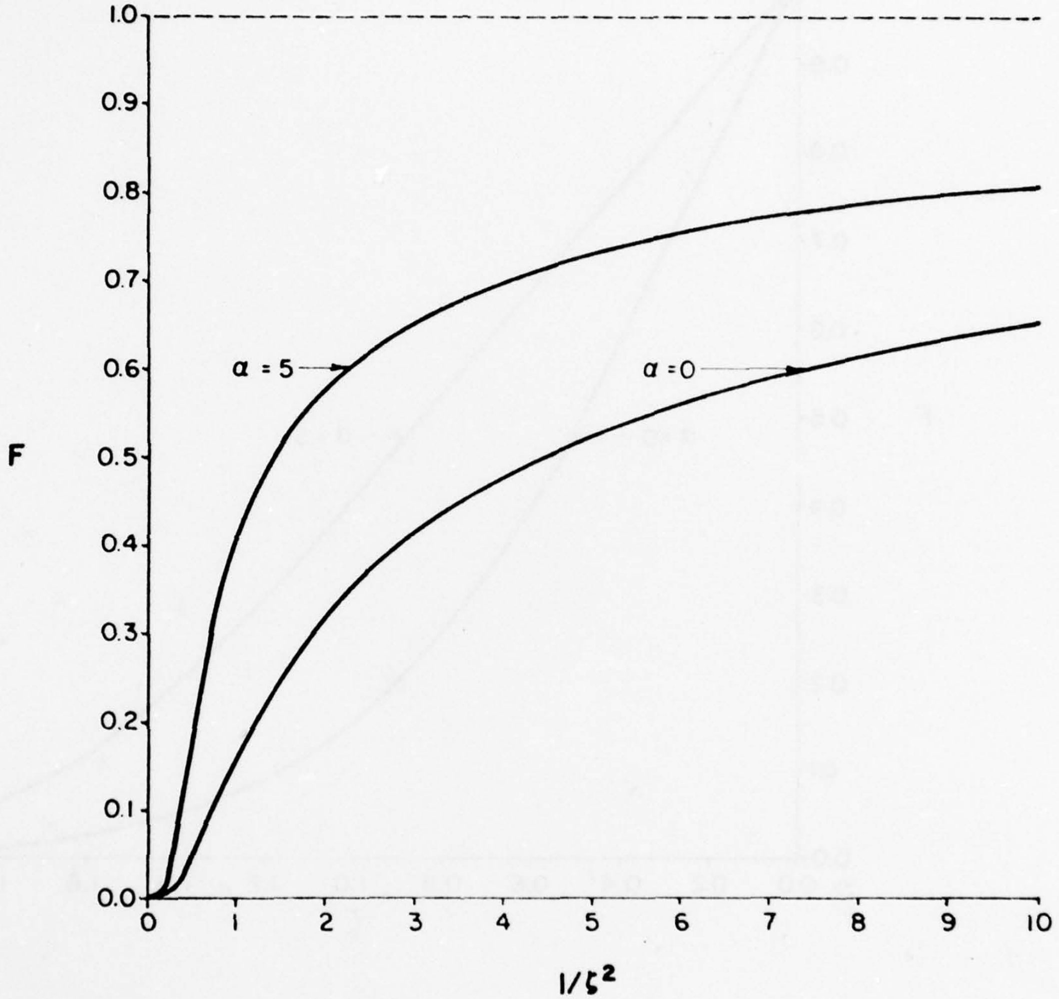


Figure 9. Normalized magnetic field function F versus ζ for α=0 and α=5. For a fixed position x, $1/\zeta^2$ is proportional to t and the above plot yields the time variation of the normalized magnetic field at the location x.

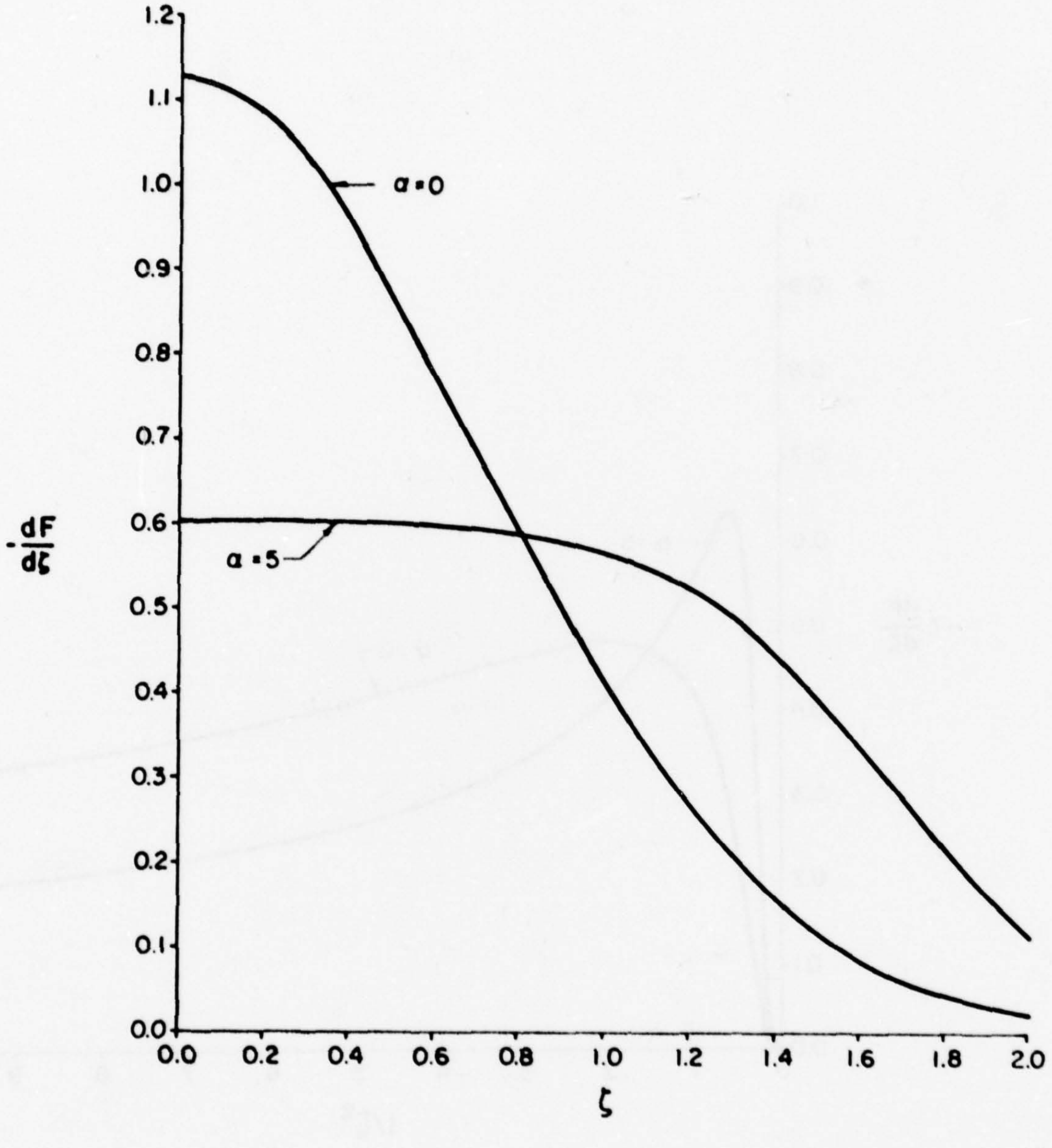


Figure 10. The slope of the normalized magnetic field function versus ζ for $\alpha=0$ and $\alpha=5$. For a fixed time, ζ is proportional to x and $\frac{dF}{d\zeta}$ is proportional to the electric field distribution at that time.

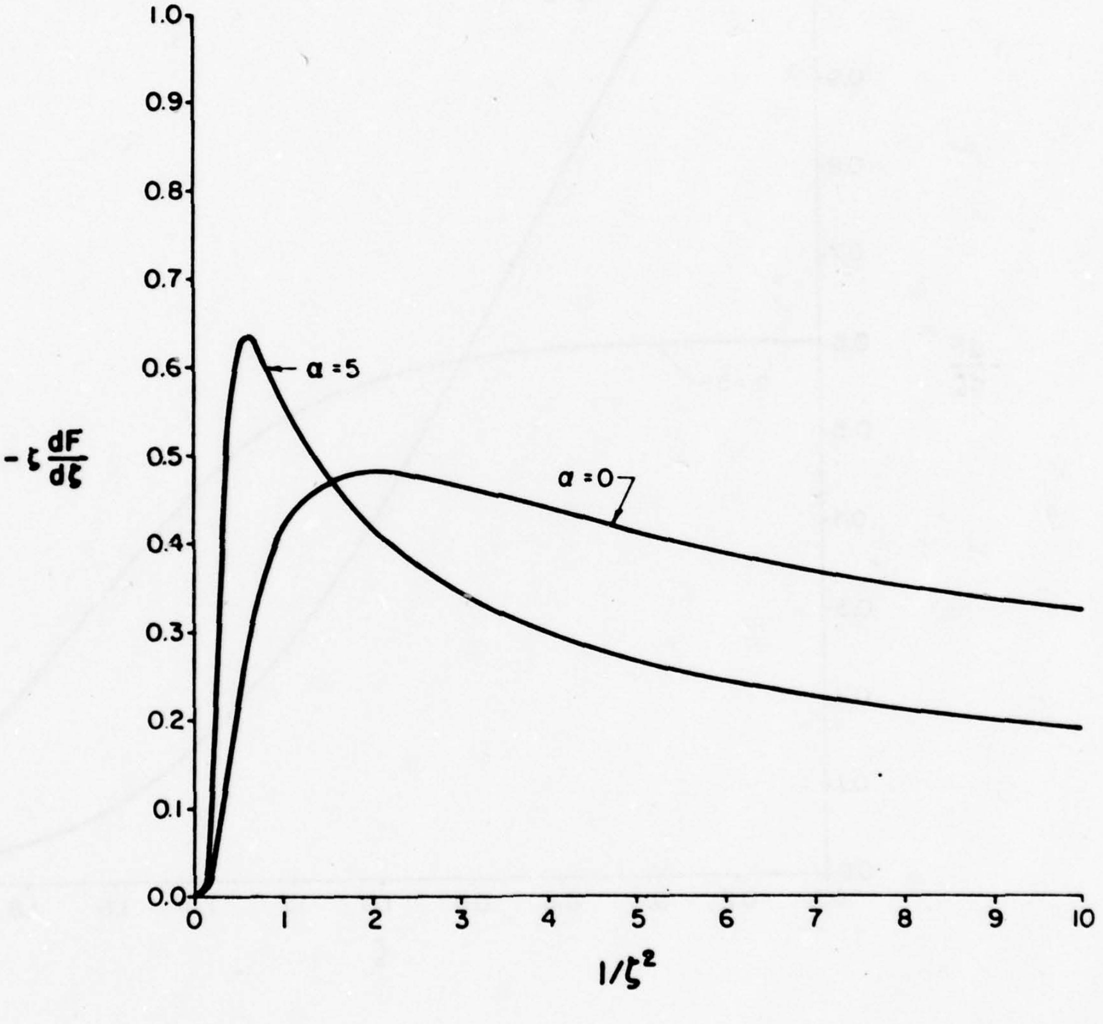


Figure 11. $\zeta\frac{dF}{d\zeta}$ versus $1/\zeta^2$ for α=0 and α=5. For a fixed location x, $\frac{1}{\zeta^2}$ is proportional to t and $\frac{dF}{d\zeta}$ indicates the time variation of the electric field at that position.

strength will never quite achieve the applied value except at the surface x = 0. For this reason it is often of interest in the study of pulse penetration phenomena to know where the field is 90% of the applied value at a given time. The time that it takes the magnetic field strength at a given location to rise to 90% of the applied value is also of interest. From Figure 8 it can be seen that the value of ζ at which F = 0.90 will have a larger value as larger values of the parameter α are considered.

The reversion of series (99) can be used effectively to calculate the value of ζ at which F = 0.90, $\zeta_{.90}$ for the value of α which occurs in the particular problem. Since (1 - F) is small in this case, only the first few terms of (99) are needed. The variation of $\zeta_{.90}$ with α is shown in Figure 12.

For a given value of the parameter α, the solution F(ζ) at some ζ = c has the same value for all combinations of x and t which are related by ζ = c. This indicates that if a certain value for the magnetic field strength has reached location x ($0 < x < \infty$) at some time t, then the same value of magnetic field strength will occur at a different location x′ = ax at some time $t' = a^2t$. For a given α, the solution F(ζ) will have a prescribed value Fp at some ζ_p. Thus, although F has a value greater than zero for $\zeta < \infty$, one can consider a "penetration thickness"

$$\delta_p \equiv \zeta_p \, (2/\sqrt{\sigma\mu_i}) \sqrt{t} \tag{101}$$

beyond which the magnetic field strength has changed by less than the fraction ′p′ of H_o. For example, the linear solution (40) has a value less than 0.01 when $\zeta > 1.82$. If it is necessary to calculate the electromagnetic fields in a plate of finite thickness, then the solutions F(ζ), dF/dζ and ζdF/dζ will be good approximations for times such that the "penetration thickness" is small with respect to the plate thickness.

CONCLUSIONS

The preceding analyses and numerical calculations demonstrate that the electromagnetic field penetration in the nonlinear case (with field dependent permeability) differs significantly from the linear case (with field independent permeability). In the linear case, the solutions for the magnetic field strength and electric field are linearly scaled by H_o and the shapes of F, dF/dζ, and

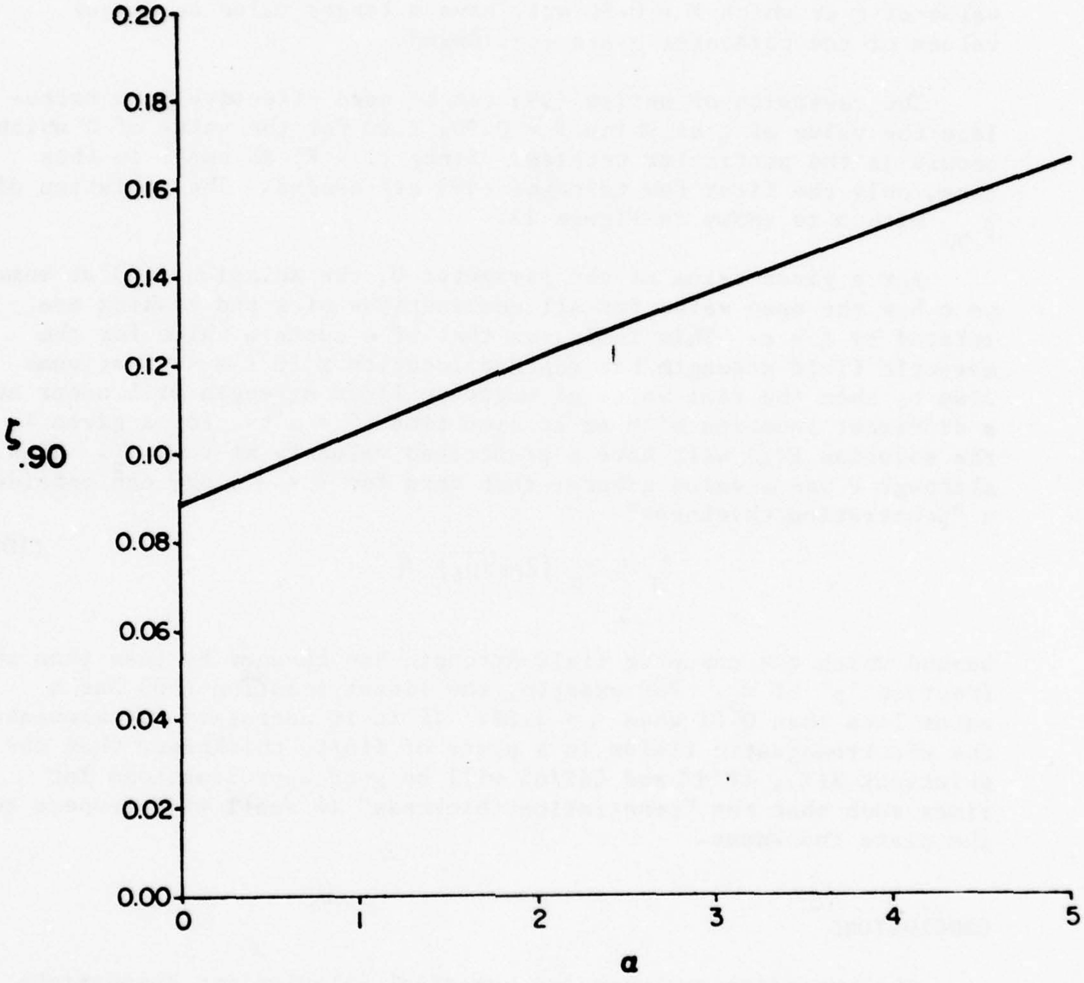


Figure 12. The variation with saturation index α of the value of ζ at which the normalized magnetic field has a value of 0.90. The relationship is almost linear for 0 ≦ α ≦ 5.

$dF/d\zeta$ remain unchanged. As might be expected, the results in the nonlinear case may not be simply scaled in the usual linear fashion. Although by the nature of the problem the final value H_o is the same, the manner in which the magnetic field strength approaches this value depends on the magnitude of H_o.

The exponentially decreasing permeability results in a more rapid penetration of the medium by the magnetic field. In other words, at a given position in the medium, the magnetic field strength H will reach a given percentage of the applied field H_o in a shorter time. Alternatively, at a given time the point at which the magnetic field strength has a prescribed value will have penetrated to a greater distance than if the permeability had remained unchanged. The more rapid penetration of the magnetic field results in a flattening of the electric field distribution at a given time. The electric field at a given position in the medium will rise more rapidly, achieve a greater maximum, and decrease more rapidly than if the permeability had remained constant. For small variations in permeability, the solutions for the electromagnetic fields differ only slightly from the solutions using a constant permeability. On the other hand, for large variations in permeability the electromagnetic field distributions vary markedly from those for the constant property case.

REFERENCE

[1] J. Crank, Mathematics of Diffusion, Oxford University Press, London, 1964, p. 166-170.

STEADY IN-PLANE DEFORMATION OF THE NONCOAXIAL PLASTIC SOIL*

Shunsuke Takagi
U. S. Army Cold Regions Research
and Engineering Laboratory
Hanover, New Hampshire 03755

ABSTRACT

This paper presents the theory of the steady in-plane plastic deformation, obeying the Coulomb yield criterion, of soils whose strain rate and stress principal directions are noncoaxial. The constitutive equations including an unknown noncoaxial angle are derived by use of the geometry of the Mohr circle and the theory of characteristics lines. A boundary value problem is solved by assigning the noncoaxial angle a set of such values that enable us to accommodate the presupposed type of flow satisfying the given boundary conditions in a given domain. The plastic material regulated by the Coulomb yield criterion in in-plane deformation is, therefore, a singular material whose constitutive equations are not constant with material but are variable with flow conditions.

*This paper is to appear in full in the "International Journal of Engineering Science".

LARGE PLASTIC DEFORMATION IN A RADIAL DRAWING PROCESS

P. C. T. Chen
U.S. Army Armament Research and Development Command
Benet Weapons Laboratory, LCWSL
Watervliet Arsenal, Watervliet, NY 12189

ABSTRACT. The problem considered is the large plastic-deformation and stresses in the flange of a radial drawing process. An analytical large strain solution is obtained on the basis of a deformation theory of anisotropic plasticity. The orthotropic sheet is rigid-plastic, isotropic in its plane and hardens according to a power law. Some numerical results are presented and discussed.

I. INTRODUCTION. In the process of deep-drawing, a thin circular blank is formed into a cylindrical cup, open at the top and closed at the base [1]. This test is considered to provide a measure of the drawability of sheet metals. The material is strain-hardening and usually anisotropic with isotropy in the plane of the sheet. The stresses in the flange while it is being drawn radially towards the throat of the die can be regarded as in a state of plane stress condition. Analytical studies have been made by Hill [1], Chiang and Kobayashi [2] Budiansky and Wang [3], Yang, [4], etc. Large strain flow theory was used in [2,3] and small strain deformation theory, in [3,4]. Recently, a finite element solution was reported by Wifi [5].

For axisymmetric plane stress problems, the earlier results [6] have shown that the differences between the flow and deformation theories are small in the small strain case. However, the differences between the two theories in the large strain case remain to be determined. In the present paper, the large strain deformation theory together with a power-hardening law was used. A new analytical solution was obtained. The numerical integration procedure can be carried out within a specified accuracy. The small-strain solution in closed form can be obtained as a special case of this new solution.

II. BASIC EQUATIONS. Suppose that the circular sheet specimen under investigation is of uniform thickness h_0 and has an initial radius b. Let the initial position of a material element in the specimen be r, and suppose that at some stage of the drawing this element has been displaced to a radial position ρ. Then the current radial displacement u of this element is

$$u = \rho - r \tag{1}$$

and the strains in the radial, circumferential, and transverse directions are given by

$$\varepsilon_r = \ln(d\rho/dr) \tag{2}$$

$$\varepsilon_\theta = \ln(\rho/r) \tag{3}$$

$$\varepsilon_z = \ln(h/h_o) \tag{4}$$

where h denotes the current thickness. All strain components are functions of r. Here we have assumed ε_z to be uniform throughout the thickness of the sheet for a given r. Since the elastic strains are assumed to be negligible and the plastic deformation is assumed to be incompressible, we have the relation

$$\varepsilon_r + \varepsilon_\theta + \varepsilon_z = 0 \tag{5}$$

The equilibrium equation in the radial direction is written in terms of the deformed coordinate as

$$(d/d\rho)(h\sigma_r) + (h/\rho)(\sigma_r - \sigma_\theta) = 0 \tag{6}$$

where σ_r and σ_θ are the stress components in the radial and circumferential directions, respectively. The other two equilibrium equations are automatically satisfied.

The total stress-strain relations derived from Hill's incremental relations [1] and used by Budiansky and Wang [3], Yang [4] have the form

$$\varepsilon_r = (\varepsilon/\sigma)(\sigma_r - \mu_p\sigma_\theta) \tag{7}$$

$$\varepsilon_\theta = (\varepsilon/\sigma)(\sigma_\theta - \mu_p\sigma_r) \tag{8}$$

where σ is the effective stress defined by

$$\sigma = (\sigma_r^2 + \sigma_\theta^2 - 2\mu_p\sigma_r\sigma_\theta)^{1/2} , \tag{9}$$

ε is the effective strain related to σ by a power law

$$\varepsilon = (\sigma/K)^m , \tag{10}$$

and μ_p, m, K are three material constants. The plastic Poisson's ratio μ_p is related to the plastic strain ratio R and the yield stress ratio ω by

$$\mu_p = R(1+R) = 1 - (2\omega^2)^{-1} \tag{11}$$

To analyze the radial drawing process in the flange, it will be necessary to solve ten equations (1) through (10) for the ten dependent variables ρ, u, h, ε_r, ε_θ, ε_z, ε, σ_r, σ_θ, and σ as functions of r and T. It will be convenient to identify the time-like variable T with the inward displacement $-U_b$ of the rim. The boundary condition at the rim is

$$\sigma_r = 0 \quad \text{at } r = b \; . \tag{12}$$

As drawing proceeds, the range of r continually decreases as material leaves the flange and enters the cup wall. This range of r is to be determined in the analysis.

III. LARGE-STRAIN SOLUTION. For convenience of analysis, the stresses are nondimensionalized and defined by

$$S_r = \sigma_r/K, \quad S_\theta = \sigma_\theta/K, \quad S = \sigma/K. \tag{13}$$

The algebraic equation (9) is satisfied identically by the following parametric representation for S_r and S_θ:

$$S_r = S\cos\phi/\sin 2\delta \tag{14}$$

$$S_\theta = S\cos(\phi+2\delta)/\sin 2\delta \tag{15}$$

where δ is the anisotropic parameter defined in the first quadrant by

$$\tan^2\delta = (1-\mu_p)/(1+\mu_p) = 1/(1+2R) = 1/(4\omega^2-1) \; . \tag{16}$$

From the evident requirements $\varepsilon_\theta \leq 0$, $S_r \geq 0$, it appears that ϕ must remain in the interval $(0, \pi/2)$.

Using the above relations (13-16), we can simplify equations (10), (7), (8) and (5) to the following forms:

$$\varepsilon = S^m \tag{17}$$

$$\varepsilon_r = S^m \sin(\phi+2\delta) \tag{18}$$

$$\varepsilon_\theta = -S^m \sin\phi \tag{19}$$

$$\varepsilon_z = -2\sin\delta \; S^m \cos(\phi+\delta) \tag{20}$$

Now all the stresses and strains are expressed in terms of S and ϕ. The thickness h can be determined by (4) and (20), and its differential form is

$$dh/h = d\varepsilon_z = 2\sin\delta\cdot S^m[\sin(\phi+\delta)d\phi - m\cos(\phi+\delta)S^{-1}dS]. \qquad (21)$$

The current position ρ can be determined by (3) and (19). Its differentiated form should be equal to that from (2) and (18). Using (2) and (3), we have the relation

$$\rho^{-1}d\rho = \exp(\varepsilon_r-\varepsilon_\theta)r^{-1}dr \qquad (22)$$

In order to solve the displacement uniquely, the strains have to satisfy the compatibility equation

$$r(d\varepsilon_\theta/dr) = \exp(\varepsilon_r-\varepsilon_\theta)-1 \qquad (23)$$

which is obtained by eliminating ρ in (2) and (3).

Substituting (18) and (19) into the above compatibility equation, we have

$$r^{-1}dr = -[\cos\phi d\phi + mS^{-1}\sin\phi dS]S^m/[\exp(f)-1] \qquad (24)$$

where

$$f = \varepsilon_r-\varepsilon_\theta = 2S^m\sin(\phi+\delta)\cos\delta \qquad (25)$$

with the aid of relations (14), (15), (21) and (22), the equation of equilibrium (6) can be reduced to

$$\rho^{-1}d\rho = [dS_r + S_r(dh/h)]/(S_\theta-S_r)$$

or

$$r^{-1}dr = e^{-f}[f_1(\phi,S)d\phi + f_2(\phi,S)S^{-1}ds] , \qquad (26)$$

where

$$f_1(\phi,S) = \frac{\sin\phi}{2\sin\delta\ \sin(\phi+\delta)} - S^m\cos\phi \qquad (27)$$

and

$$f_2(\phi,S) = \frac{-\cos\phi}{2\sin\delta\ \sin(\phi+\delta)} + \frac{mS^m\ \cos\phi\cos(\phi+\delta)}{\sin(\phi+\delta)} \tag{28}$$

The three ordinary differential equations (22), (24) and (26) form a system and can be used to solve for S, ϕ and ρ as functions of r and T. It will be convenient to identify the time-like variable T with the inward displacement $-U_b$ of the rim. With the aid of equations (1), (3), (14), and (19), the boundary condition (12) determines the values of S, ϕ and ρ at r = b,

$$\phi_b = \pi/2, \quad S_b = [\ln(b/\rho)]^{\frac{1}{m}}, \quad \rho_b = b-|U_b| \tag{29}$$

For the radial drawing problem considered in this paper, it was found advantageous to solve S and ρ as functions of ϕ. Elimination of the space variable r between (24) and (26) gives

$$S^{-1}(dS/d\phi) = g(S,\phi) \tag{30}$$

where

$$g(S,\phi) = \frac{\tan\phi + \tan\delta[fe^{-f}/(1-e^{-f})]}{1-m\tan\phi\tan\delta[f/(1-e^{-f})]-2mS^m\sin\delta\cos(\phi+\delta)} \tag{31}$$

With the aid of relations (22) and (30), equation (24) becomes

$$\rho^{-1}(d\rho/d\phi) = -S^m[\cos\phi+m\sin\phi\cdot g(\phi,S)]/(1-e^{-f}) \ .$$

The two ordinary differential equations (30) and (32) with the initial condition (29) form a system for solving S and ρ as functions of ϕ. The system can be solved numerically using a fourth-order Runge-Kutta integration process [7]. The numerical integration procedure is to be carried out until the value of ρ reaches ρ_1 = a at the die throat (junction between flange and cup).

IV. SMALL-STRAIN SOLUTION. The governing equations for large-strain solution can be reduced to the special case for small-strain solution. Note that for small strain

$$\varepsilon = S^m << 1, \quad f = 2\varepsilon\cos\delta\sin(\phi+\delta) << 1,$$
$$S^m/(e^f-1) \simeq \varepsilon/f = [2\cos\delta\sin(\phi+\delta)]^{-1}. \tag{33}$$

Therefore the governing equations (24) and (26) for large strain solution can be simplified to

$$r^{-1}dr = - [\sin 2\delta(\cot\delta + \cot\phi)]^{-1}(\cot\phi d\phi + mS^{-1}dS) \tag{34}$$

and

$$r^{-1}dr = [\sin 2\delta(\tan\delta + \tan\phi)]^{-1}(\tan\phi d\phi - S^{-1}dS) \tag{35}$$

which are identical to those in [6] obtained from the basic equations for small strain only. These two ordinary differential equations can be integrated explicitly with the boundary values S_b and ϕ_b at $r = b$. The results can be presented in the following form [8]:

$$S/S_b = G(\phi), \tag{36}$$

and

$$(b/r)^2 = F(\phi), \tag{37}$$

where

$$G(\phi) = (\sin\phi - m^{-1}\cot\delta\cos\phi)^{-\mu_1}\exp[\frac{(m-1)\cot\delta}{m^2 + \cot^2\delta}(\frac{\pi}{2} - \phi)] \tag{38}$$

$$F(\phi) = \frac{\sin(\phi+\delta)/\cos\delta}{(\sin\phi - m^{-1}\cot\delta\cos\phi)^{\mu_2}}\exp[\frac{(m^2-1)\cot\delta}{m^2 + \cot^2\delta}(\frac{\pi}{2} - \phi)] \tag{39}$$

$$\mu_1 = (m + \cot^2\delta)/(m^2 + \cot^2\delta), \tag{40}$$

and

$$\mu_2 = m\csc^2\delta/(m^2 + \cot^2\delta). \tag{41}$$

The range of r is known to extend from the die throat $r = a$ to the rim $r = b$. Thus the domain of the parameter ϕ is $\phi_1 \leq \phi \leq \pi/2$ where ϕ_1 satisfies

$$(b/a)^2 = F(\phi_1) . \tag{42}$$

Now the small strain solution for the displacement strains and stresses at any location can be calculated.

V. NUMERICAL RESULTS AND DISCUSSIONS. The small strain solution can offer information on the asymptotic behavior at the initial stage of drawing process as discussed in [4]. In this section, the numerical results and discussions based on large strain solution are presented. The governing equations (30) and (32) with condition (29) for large strain solution form a system for solving S and ρ as functions of ϕ. The system can be solved numerically using a fourth-order Runge-Kutta integration process [7]. The numerical integration procedure is to be carried out until the value of ρ reaches $\rho_1 = a$ at the die throat (junction between flange and cup). In order to reach the limit state within a specified accuracy, say 0.01%, an iterative approach is used. The results corresponding to this limit state $\rho_1/a = 1$ are denoted by S_1, ϕ_1, r_1, u_1, h_1, $(\varepsilon_r)_1$, $(\varepsilon_\theta)_1$, $(\varepsilon_z)_1$, $(S_r)_1$, and $(S_\theta)_1$. The quantity r_1 represents the initial position for the deformed particle at the die throat. The values for r_1, S_1, ϕ_1, etc. are functions of the inward rim displacement $|U_b|$. As drawing proceeds, the value for r_1 increases and the range of r in the flange ($r_1 \leq r \leq b$) continually decreases as material leaves the flange and enters the cup wall. Numerical results corresponding to these limiting states are presented in Figures 1 to 3 for the case $R = 1$, $m = 5$ and $b/a = 2$. Figure 1 shows the radial displacement U_1 and thickness h_1 at the die throat as functions of inward rim displacement $|U_b|$. We can also see clearly the limiting value r_1 as a function of $|U_b|$ since $r_1 = a + |u_1|$. As drawing proceeds with increasing $|U_b|$, the values for r_1, $|u_1|$, h_1 all increase. Figure 2 shows the strains $(\varepsilon_r)_1$, $(\varepsilon_\theta)_1$ and $(\varepsilon_z)_1$, at the die throat as functions of $|U_b|$. The maximum value for $(\varepsilon_r)_1$ occurs at $|U_b/a| \approx .65$. Figure 3 shows the stresses $(S_r)_1$ and $(S_\theta)_1$ at the die throat as functions of $|U_b|$. The maximum drawing stress $(\sigma_r)_1 \approx 0.478K$ occurs at $U_b \simeq -0.20a$ and the maximum drawing force $(h\sigma_r)_1 \approx 0.474\ h_oK$ occurs at $U_b \approx -0.22a$. This result is in excellent agreement with that by Budiansky and Wang [3] based on flow theory using a finite difference approach. A more detailed comparison of the large strain solution based on different theories remains to be done.

In order to demonstrate the results for spatial distributions at various stages of drawing process, we include Figures 4 and 5. Figure 4 illustrates the deformed shapes of the flanges for $-U_b = 0.1$, 0.2, 0.3 and 0.4. We can see how the flange deforms as drawing proceeds. Figure 5 shows the stresses σ_r/K and σ_θ/K in the flange at several stages of drawing process with $-U_b =$ 0.1, 0.2 and 0.4.

REFERENCES.

1. Hill, R., Mathematical Theory of Plasticity, Oxford University Press, 1950, Chapter 12.

2. Chiang, D. C. and Kobayashi, S., "The Effect of Anisotropy and Work-Hardening Characteristics on the Stress and Strain Distribution in Deep Drawing," Journal of Engineering for Industry, Vol. 88, 1966, pp. 443-448.

3. Budiansky, B., and Wang, N. M., "On the Swift Cup Test," Journal of Mechanics and Physics of Solids, Vol. 14, 1966, pp. 357-374.

4. Wei Hsuin Yang, "Axisymmetric Plane Stress Problems in Anisotropic Plasticity," Journal of Applied Mechanics, Vol. 36, 1969, pp. 7-14.

5. Wifi, A. S., "An Incremental Complete Solution of the Stretch-Forming and Deep-Drawing of a Circular Blank Using a Hemispherical Punch," Int. J. Mech. Sci., Vol. 18, 1976, pp. 23-31.

6. Chen, P. C. T., "Elastic-Plastic Analysis of a Radially Stressed Annular Plate," Journal of Applied Mechanics, Vol. 44, 1977, pp. 167-169.

7. IBM System/360 Scientific Subroutine Package (360A-CM-03X) Version III Programmers Manual, 4th Edition, 1968, pp. 333-336.

8. Chen, P. C. T., "Fully Plastic Deformation in Anistropic Annular Plate Under Internal Pressure," Transactions of the Twenty-Third Conference of Army Mathematicians, ARO Report 78-1, 1978, pp. 105-120.

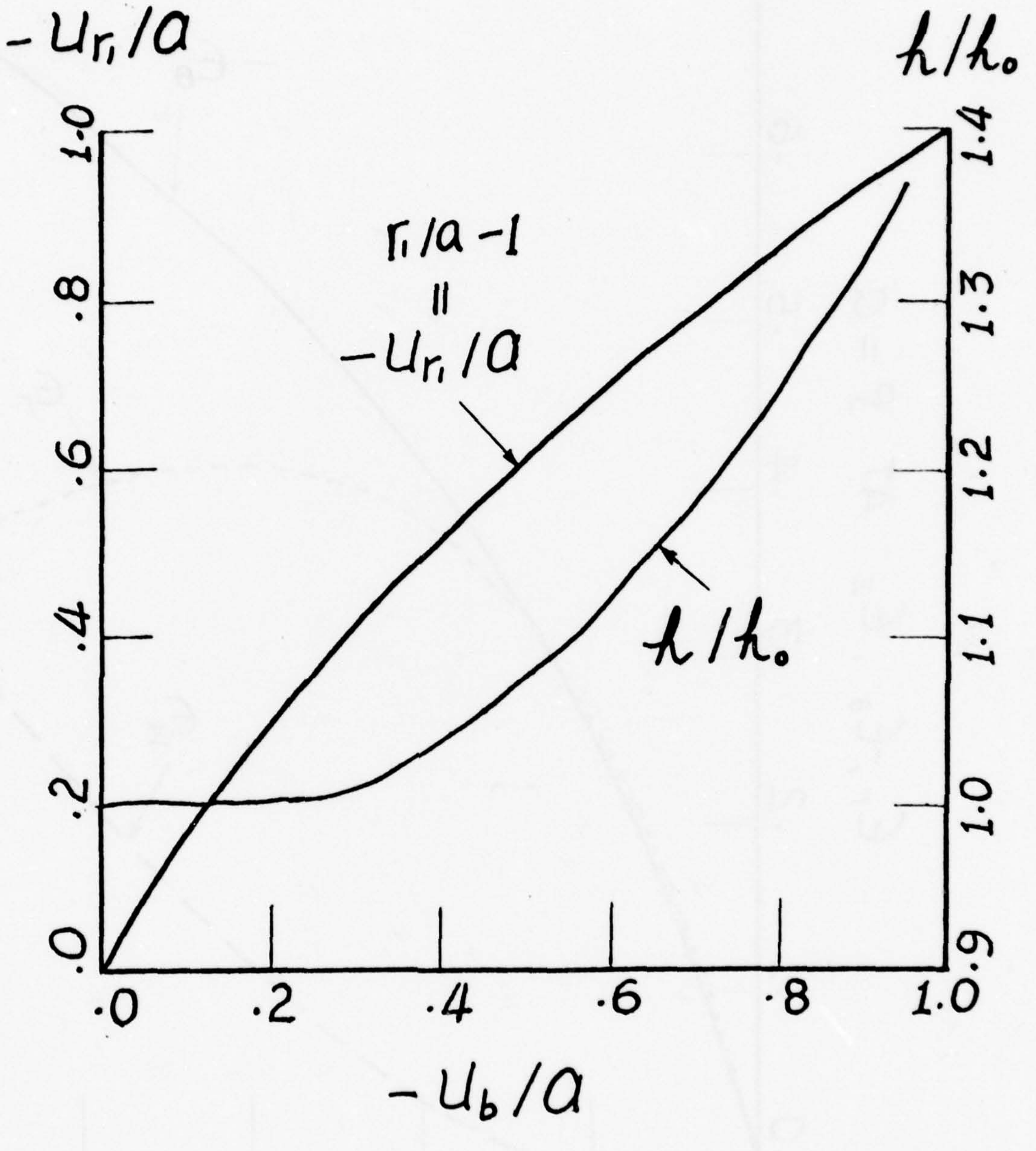


Figure 1. Radial displacement and thickness at the die throat as functions of inward rim displacement

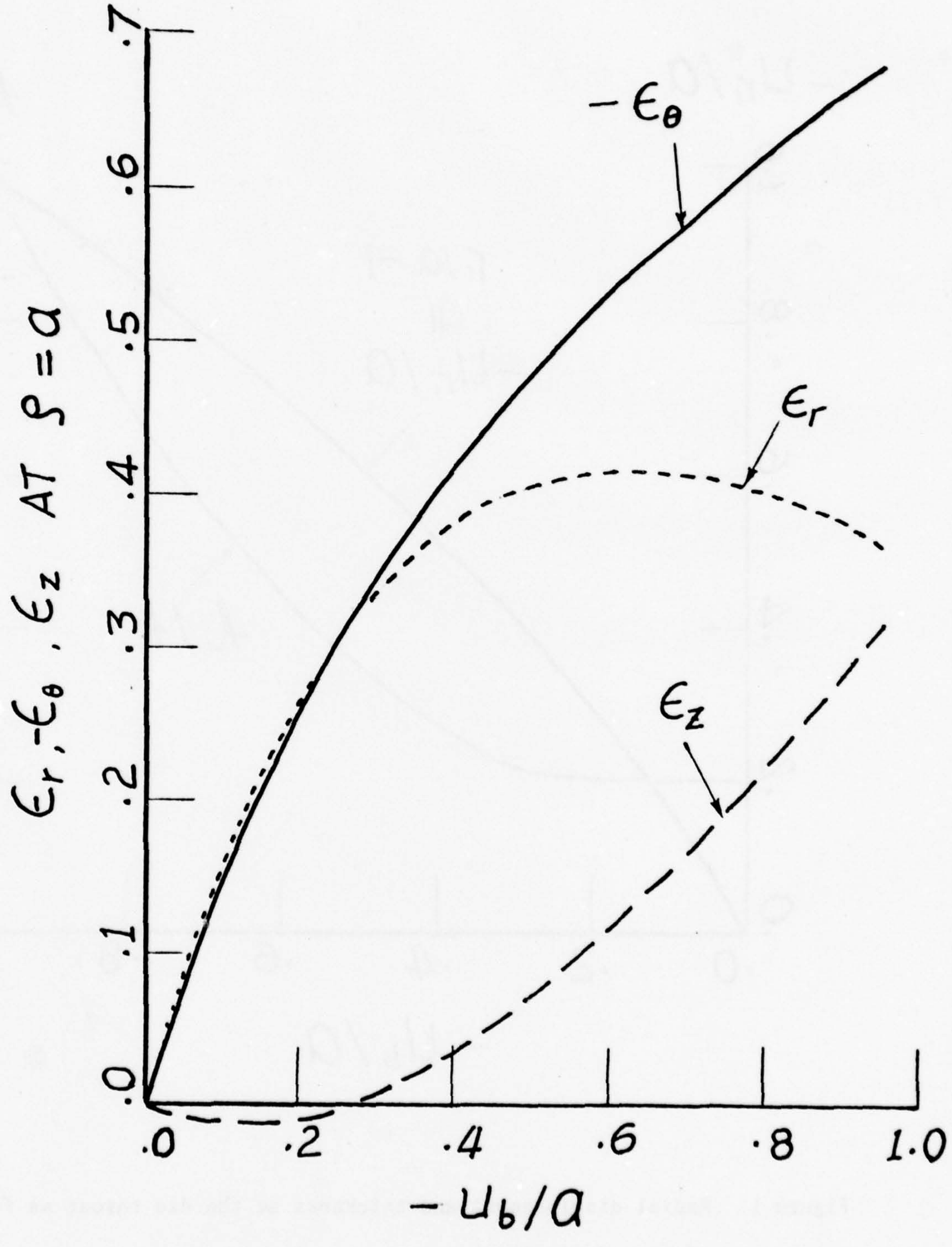


Figure 2. The strains at the die throat as functions of inward rim displacement

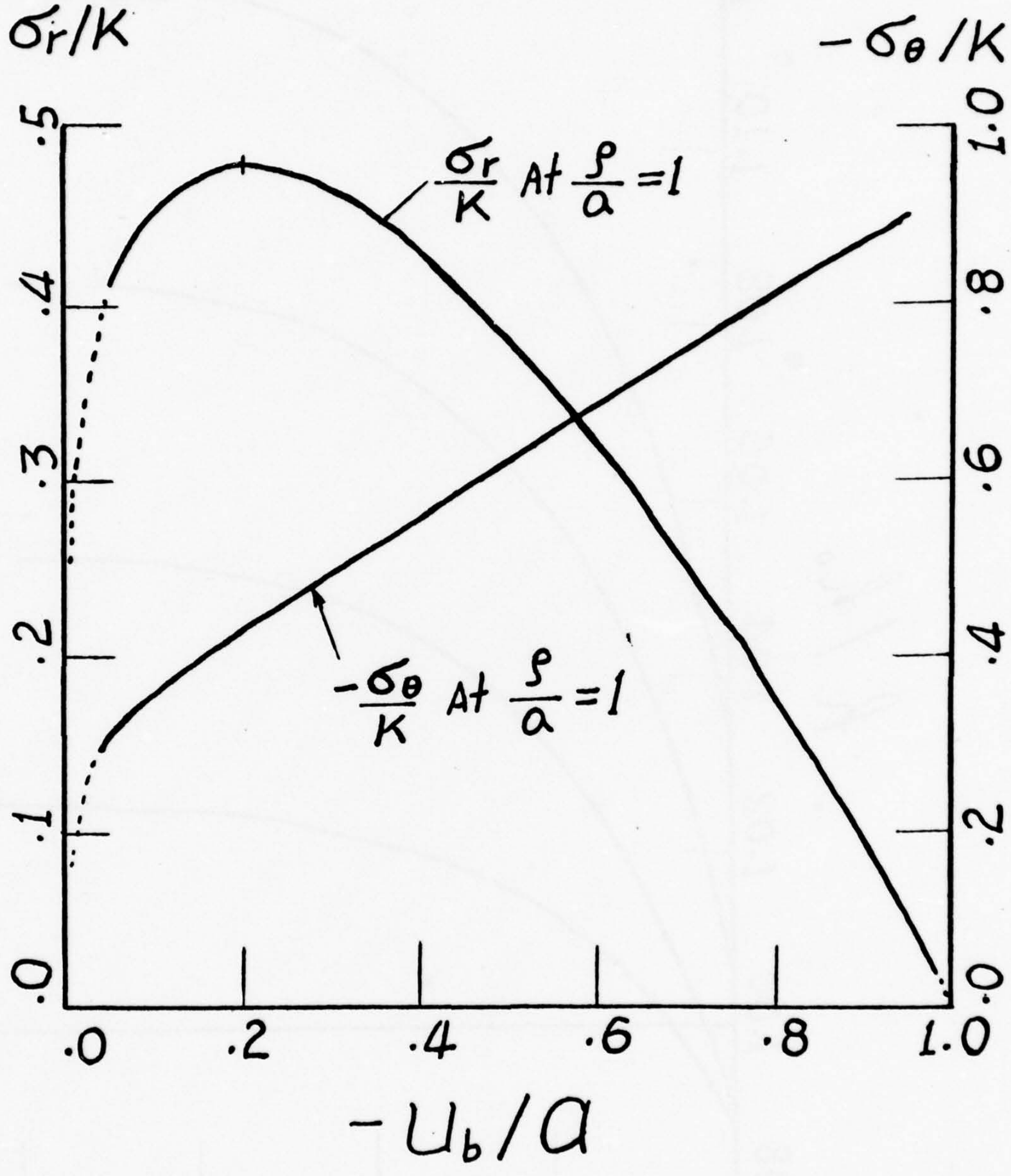


Figure 3. The stresses at the die throat as functions of inward rim displacement

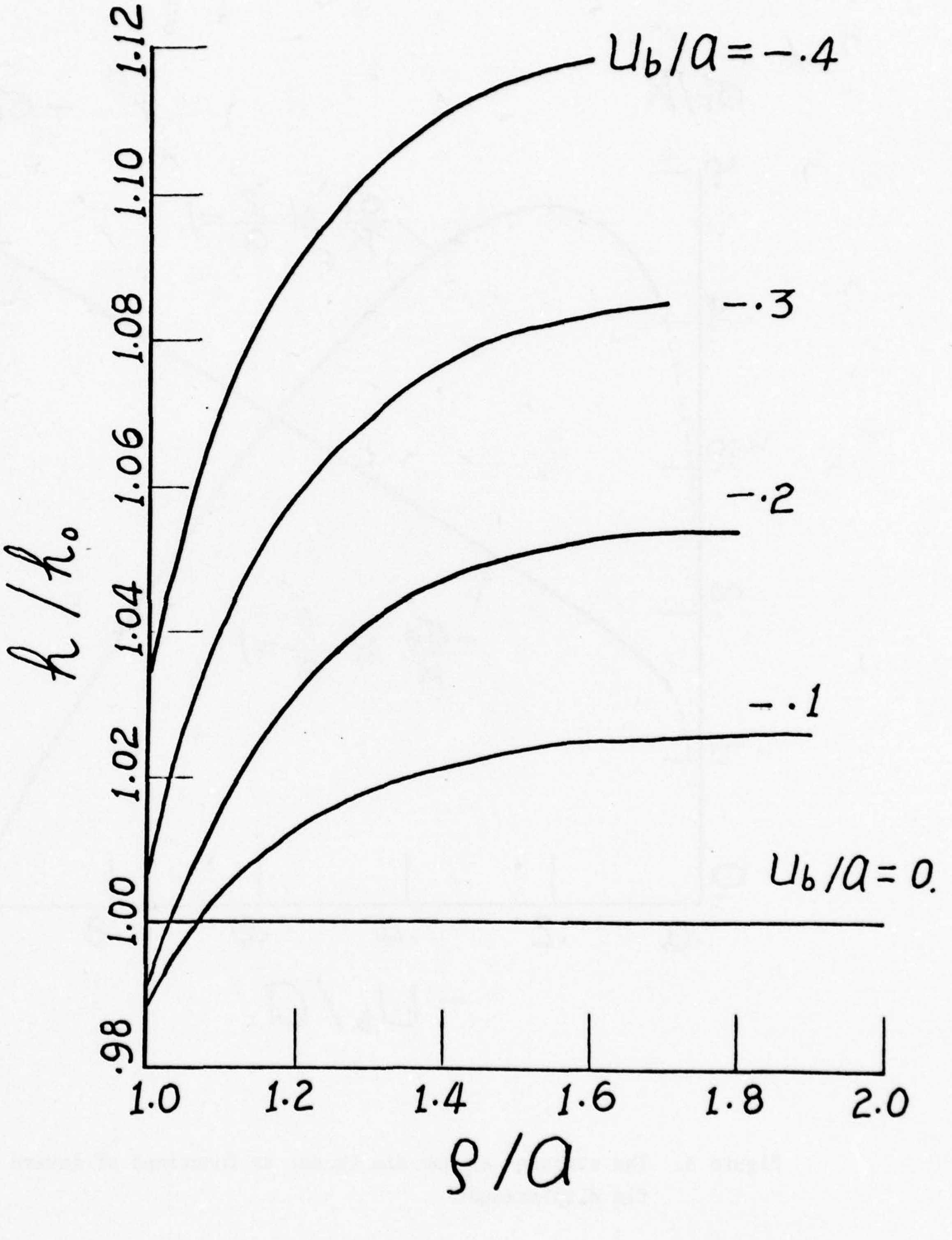


Figure 4. The deformed shapes of the flange as drawing proceeds

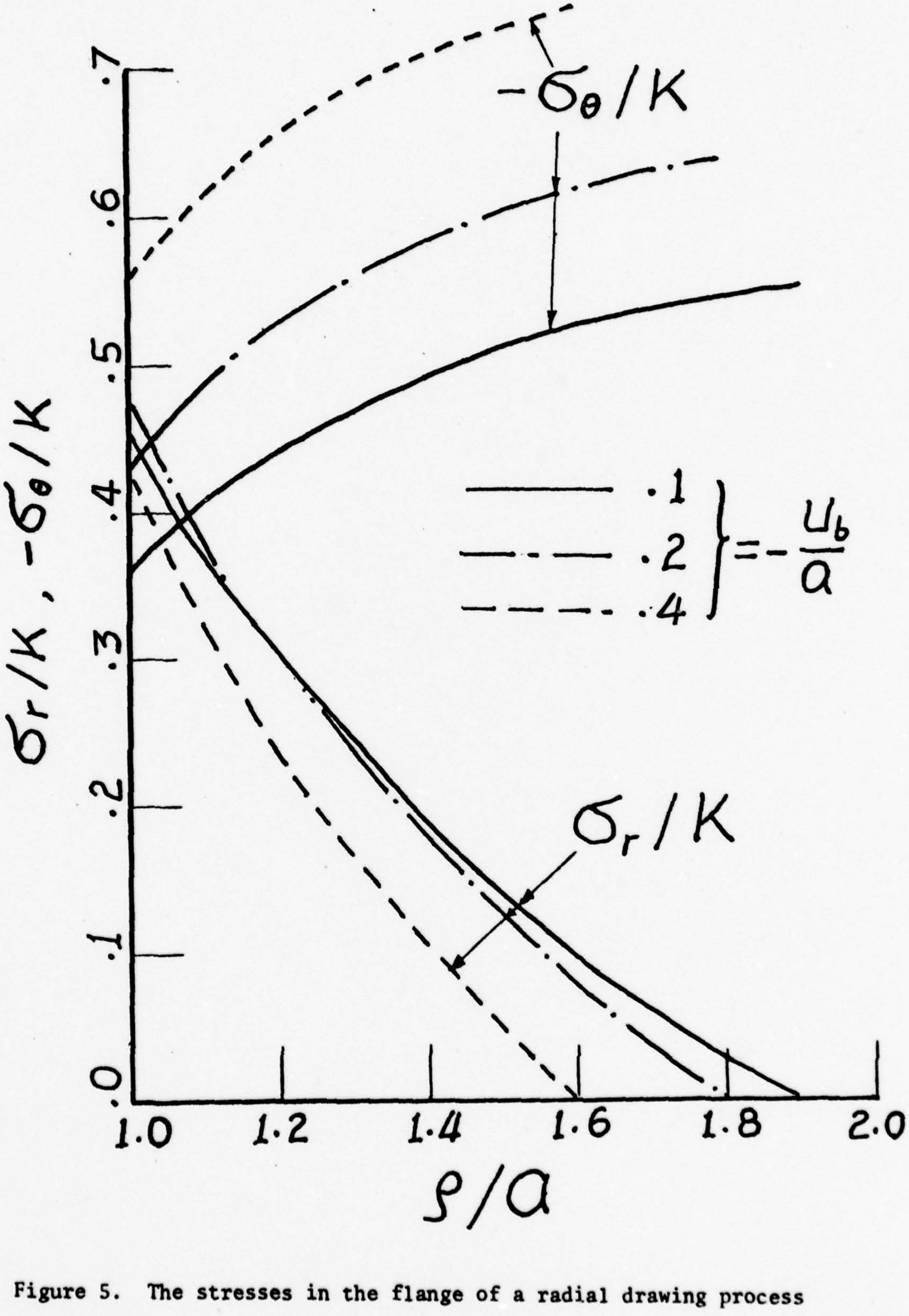


Figure 5. The stresses in the flange of a radial drawing process

MAJORIZATION FORMULAS FOR A BIHARMONIC FUNCTION OF TWO VARIABLES

J. Barkley Rosser
Mathematics Research Center
University of Wisconsin-Madison
610 Walnut Street
Madison, Wisconsin 53706

ABSTRACT. Biharmonic functions are much used in the theory of elastic solids. A problem of long standing has been to produce a formula involving only the boundary conditions which gives a bound for the value of a biharmonic function in the interior of a region. Using results of Miranda, this problem is solved if the region is a circle or rectangle.

1. BACKGROUND. We propose to extend majorization formulas given by Miranda in [2]. He was considering a biharmonic function $u(x,y)$; that is,

$$\nabla^2\nabla^2 u = 0 \tag{1.1}$$

in a region T. He allowed T to be multiply connected, but required the curvature of the boundary (or boundaries) to be continuous around the boundary, FT, and the tangent to be continuously turning. On the boundary, values of u were specified,

$$u = f , \tag{1.2}$$

and values of the inward normal derivative were specified,

$$\frac{du}{d\nu} = g . \tag{1.3}$$

He considered f and g as functions of the arc length around the boundary. He also assumed that u and its first partial derivatives are continuous in T, up to and including (one-sided) continuity at the boundary.

His Theorem II says that if $f \equiv 0$, then in T

$$|u(x,y)| \leq \sqrt{2\phi(x,y)}\, \max|g| , \tag{1.4}$$

where the maximum of $|g|$ is taken over the boundary and $\phi(x,y)$ is the function which satisfies

$$\nabla^2\phi = -1 \tag{1.5}$$

inside T and has $\phi \equiv 0$ on FT.

Sponsored by the United States Army under Contract No. DAAG29-75-C-0024.

Upon relaxing the requirement that u be zero on the boundary, Miranda also had a majorization formula. He assumed continuity of f' and g, and concluded that there are nonnegative constants K_1 and K_2, depending solely on the region T, such that in T one has

$$|u(x,y)| \leq K_1\delta[\max|g| + \max|f'|] + (1 + K_2\delta)\max|f| \ , \tag{1.6}$$

where δ is the (minimum) distance from (x,y) to FT. See his Theorem VI.

This has the shortcoming that no clue is available as to what might be the sizes of K_1 and K_2.

Professor L. Collatz has pointed out in conversation that, if one is given boundary conditions for u, it may be possible to contrive a specifically given $\bar{u}$ whose boundary conditions are not greatly different from those of u. Then, if one had specific constants in (1.6), one could bound the difference between u and $\bar{u}$ by means of (1.6). We undertake to find in two cases majorization formulas involving specific constants that can be used as Professor Collatz suggests.

We will consider the two cases where T is a circle and where T is a rectangle, and will supply majorization formulas with specific constants; for the rectangle we have to assume in addition that f'' is of bounded variation, and will find $\max|f''|$ appearing in the corresponding majorization formula. We also need a slight additional hypothesis on u at the four corners of the rectangle.

In Section 2, we collect some auxiliary formulas. In Sections 3 and 4 respectively, we treat the circle and rectangle. The relevant majorization formulas are given near the ends of the two sections in Theorems 1 and 2. In Section 5 we discuss the possibility of weakening the hypotheses of Theorem 2.

2. AUXILIARY FORMULAS.

We collect here various pieces of information that will be useful in subsequent sections.

Let $z = re^{i\theta}$ be a complex variable. If $|z| < 1$, then

$$\frac{1}{1 - z} = 1 + z + z^2 + \cdots \ . \tag{2.1}$$

Integrating gives

$$\ell n(1 - z) = -z - \frac{z^2}{2} - \frac{z^3}{3} - \cdots \ . \tag{2.2}$$

Taking real and imaginary parts gives

$$\frac{1}{2}\ \ell n(1 - 2r\cos\theta + r^2) = -\sum_{n=1}^{\infty} \frac{r^n \cos n\theta}{n} \ , \tag{2.3}$$

$$\arctan \frac{r\sin\theta}{1 - r\cos\theta} = \sum_{n=1}^{\infty} \frac{r^n \sin n\theta}{n} \ . \tag{2.4}$$

Taking $r = 1$ in (2.3) gives

(2.5) $$\frac{1}{2}\ln 2 + \frac{1}{2}\ln(1 - \cos\theta) = -\sum_{n=1}^{\infty} \frac{\cos n\theta}{n} .$$

Then

(2.6) $$\int_0^{\frac{\pi}{2}} \ln(1 - \cos\theta)d\theta = -\frac{\pi}{2}\ln 2 - 2C ,$$

where C is Catalan's constant,

(2.7) $$C = \sum_{k=0}^{\infty} (-1)^k(2k + 1)^{-2} .$$

Properties of C are discussed on p. 807 of Abramowitz and Stegun [1], and a value to 18 decimals is given on p. 812. Rounded off to 5 decimals, it is

(2.8) $$C \cong 0.91597.$$

From (2.5), we get

(2.9) $$\int_{\frac{\pi}{2}}^{\pi} \ln(1 - \cos\theta)d\theta = 2C - \frac{\pi}{2}\ln 2 .$$

Taking $r = 1$ in (2.4) gives

(2.10) $$\frac{\pi}{2} - \frac{\theta}{2} = \sum_{n=1}^{\infty} \frac{\sin n\theta}{n} \qquad (0 < \theta < 2\pi) .$$

Taking $r = -1$ in (2.4) gives

(2.11) $$\frac{\theta}{2} = \sum_{n=1}^{\infty} \frac{(-1)^{n+1}\sin n\theta}{n} \qquad (-\pi < \theta < \pi) .$$

By contour integration we can show that for $0 \leq r < R$

(2.12) $$\frac{1}{2\pi}\int_0^{2\pi} \frac{R^2 - r^2}{\ell^2} dt = 1$$

(2.13) $$\frac{(R^2 - r^2)^2}{2\pi R}\int_0^{2\pi} \frac{R - r\cos(t - \theta)}{\ell^4} dt = 1 ,$$

where

$$\ell^2 = R^2 - 2Rr\cos(t - \theta) + r^2 \, . \tag{2.14}$$

Indeed (2.12) is worked out on pp. 112-113 of Whittaker and Watson [3] as an illustrative example of contour integration.

3. FORMULA FOR A CIRCLE. Let us take Miranda's region T to be a circle. Let u satisfy (1.1), (1.2), and (1.3). Let the circle have radius R. Choose polar coordinates r and θ, with the origin at the center of the circle. Then, by a formula of Lauricella [4], we have

$$\begin{aligned} u(r,\theta) = {} & \frac{1}{2\pi} \int_0^{2\pi} f(t) \, \frac{R^2 - r^2}{\ell^2} \, dt \\ & - \frac{r(R^2 - r^2)}{2\pi R} \int_0^{2\pi} f'(t) \, \frac{\sin(t - \theta)}{\ell^2} \, dt \\ & + \frac{R^2 - r^2}{4\pi R} \int_0^{2\pi} g(t) \, \frac{R^2 - r^2}{\ell^2} \, dt \, , \end{aligned} \tag{3.1}$$

where ℓ is the distance from the point (r,θ) to the point (R,t); that is, ℓ^2 is given by (2.14).

This formula is usually cited with a minus before the last term on the right, because Lauricella was taking $g(\theta)$ to be $\partial u/\partial r$, whereas we are taking $g(\theta)$ to be the inward normal derivative. Integration by parts in the second term on the right of (3.1) reduces (3.1) to

$$u(r,\theta) = \frac{(R^2 - r^2)^2}{2\pi R} \int_0^{2\pi} f(t) \, \frac{R - r\cos(t - \theta)}{\ell^4} \, dt + \frac{R^2 - r^2}{4\pi R} \int_0^{2\pi} g(t) \, \frac{R^2 - r^2}{\ell^2} \, dt \, . \tag{3.2}$$

This formula must be used with some caution. If, at $\theta = \theta_0$, $f'(\theta)$ has a jump discontinuity and $g(\theta)$ is continuous, then the limit of the inward normal derivative as one approaches the circumference along the ray $\theta = \theta_0$ is

$$g(\theta_0) + \frac{f'(\theta_0+) - f'(\theta_0-)}{\pi} \, . \tag{3.3}$$

This and other idiosyncracies of (3.2) are discussed in Picone [5]; see especially pp. 216-217 and pp. 257-261.

If we make the same assumptions that Miranda made to derive (1.4) and (1.6), namely that u and its first partial derivatives are continuous in T up to and including the circumference, that will assure us that f' and g are continuous. Then (3.2) defines the one and only u satisfying (1.1), (1.2), and (1.3) for the circle.

As the integrands in (2.12) and (2.13) are positive, one can immediately conclude from (3.2) by the mean value theorem that, for some $\tau(r,\theta)$ and $\sigma(r,\theta)$,

$$u = f(\tau(r,\theta)) + \frac{R^2 - r^2}{2R} g(\sigma(r,\theta)) \ . \tag{3.4}$$

As a consequence, if $g(\sigma)$ is nonpositive for all σ, we have

$$u \leq \max f \ . \tag{3.5}$$

Similarly, if $g(\sigma)$ is nonnegative for all σ, one has

$$u \geq \min f \ . \tag{3.6}$$

In Picone [5], at the bottom of p. 216, these surprising conclusions are attributed to Miranda, who apparently used essentially the same proof.

From (3.4), we immediately get our majorization formula.

Theorem 1. Let u be a biharmonic function satisfying (1.1), (1.2), and (1.3) in a circle of radius R, such that u and its first partial derivatives are continuous up to and including the circumference. Then at a point r units from the center,

$$|u| \leq \max|f| + \frac{R^2 - r^2}{2R} \max|g| \ . \tag{3.7}$$

Clearly this is much superior to Miranda's Theorem VI (see (1.6)). Also, if $f \equiv 0$ it is appreciably better than Miranda's Theorem II (see (1.4)). Indeed, for the region under consideration, the ϕ of Miranda's Theorem II is

$$\phi = (R^2 - r^2)/4 \ . \tag{3.8}$$

This ϕ is biharmonic. If we put it for u in (3.7), we find that it itself is given as an upper bound for itself. Clearly this is the best possible. However, if we use this ϕ for u in Miranda's Theorem II, namely (1.4), a bound for u at the center of the circle is given which is $\sqrt{2}$ times as great as the actual value of u at that point.

4. FORMULA FOR A RECTANGLE. If T is a simply connected region more irregular than a circle, one may map it conformally into a circle. If the region has a smooth enough boundary, the conformality may extend out to the boundary, so that normal derivatives go into normal derivatives. If one has a formula for the conformal mapping, it may be tractable enough that one can calculate factors of proportionality for the various derivatives. Thus one can sometimes convert (3.7) into a majorization formula for the more general region.

When one maps a rectangle into a circle, conformality certainly does not extend out to the boundary at the four corners. In any case the formula for the transformation is much too complicated to be of much use. So we give a separate treatment for the case that T is a rectangle.

Let a and b be the lengths of the sides of the rectangle. Choose coordinates so that one corner of the rectangle is at the origin, and the rectangle extends a units along the positive x-axis, and b units along the positive y-axis.

Let u satisfy (1.1), (1.2), and (1.3). We assume that u and its first partial derivatives are continuous up to and including the perimeter. Also, on each side we assume that f" is of bounded variation in the closed interval consisting of that siue.

At the corners, strange things can happen. For one thing, a normal derivative is not defined at a corner, nor do the normal derivatives along the two sides have to have the same limits as one approaches a corner. However, continuity of $\partial u/\partial x$ as one goes to the corner means that the limit of a normal derivative on a vertical side as one approaches an upper corner must be either f' or -f' along the top at the corner. Whether it is f' or -f' depends on which direction is taken as increasing arc length, and will be different at different corners. We assume further:

Boundedness Hypothesis. For each corner there is a neighborhood of that corner within which $\nabla^2 u$ is bounded.

As an illustration of a need for a boundedness hypothesis, we cite the following example. Consider the first quadrant of a circle of radius unity with center at the origin. If a function is harmonic inside this region and is zero around the boundary, it must be identically zero by the maximum principle. But note the harmonic function

$$v = r^2 \sin 2\theta - \frac{\sin 2\theta}{r^2} \quad . \tag{4.1}$$

It is zero around the boundary, except for an indeterminacy at the origin, where r = 0. However, v is certainly not identically zero. It is because of the unboundedness at the origin that the usual maximum principle for harmonic functions fails. The reason we invoke our Boundedness Hypothesis is to avert a similar difficulty.

As we said, we let u satisfy (1.1), (1.2), and (1.3). Choose $\bar{u}$ to be the harmonic function inside the rectangle such that on the perimeter

$$\bar{u} = f \ . \tag{4.2}$$

As $\bar{u}$ is harmonic, it is biharmonic. So $u - \bar{u}$ is biharmonic, and is zero on the perimeter. Applying Miranda's Theorem II (see (1.4)), we get

$$|u - \bar{u}| \le \sqrt{2\phi}\, \max|g - \bar{g}| \ , \tag{4.3}$$

where $\bar{g}$ is the value of the inward normal derivative for $\bar{u}$, and ϕ is as in (1.4).

We will later show that $\bar{u}$ and its first partial derivatives are continuous up to and including the perimeter. So $u - \bar{u}$ satisfies those of Miranda's hypotheses. However, Miranda also assumed that his region had a boundary with continuous curvature and continuously turning tangent. Lacking these, we proceed as follows.

On p. 99 of Miranda [2] is proved the Lemma that $u\nabla^2 u$ is continuous in the region, and approaches zero on the boundary, if $f \equiv 0$. One can easily modify Miranda's proof to show that if some segment of the boundary has continuous curvature and a continuously turning tangent, then within a closed portion of that segment $u\nabla^2 u$ approaches zero uniformly as one approaches the boundary. So in the interior of each side of the rectangle, one has $(u - \bar{u})\nabla^2(u - \bar{u})$ approaching zero as one approaches the perimeter. In the neighborhood of a corner, $\nabla^2 u$ is bounded. But $\bar{u}$ is harmonic, so that $\nabla^2\bar{u} = 0$. Hence $\nabla^2(u - \bar{u})$ is bounded. But $u - \bar{u}$ approaches zero, continuously. So we conclude that $(u - \bar{u})\nabla^2(u - \bar{u})$ approaches zero as one approaches a corner.

So Miranda's Lemma holds for the rectangle, and the rest of the proof proceeds just as in Miranda [2].

For the ϕ of (4.3), we may start with

$$\phi^* = \frac{1}{2}x(a - x) . \tag{4.4}$$

This is zero on the two vertical sides of the rectangle. Now add to ϕ^* a harmonic function ϕ^{**} which is zero on the two vertical sides, and equal to

$$-\frac{1}{2}x(a - x)$$

on the top and bottom. Then $\phi = \phi^* + \phi^{**}$. By the principle of the maximum, ϕ^{**} will be everywhere nonpositive. So the ϕ of (4.3) will be bounded above by (4.4). It will also be nonnegative, because it is zero on the perimeter. By a similar argument, ϕ is bounded above by

$$\phi^{***} = \frac{1}{2}y(b - y) .$$

We have

$$|u| \le |u - \bar{u}| + |\bar{u}| .$$

But, by the principle of the maximum,

$$|\bar{u}| \le \max|f| .$$

So, by (4.3) we get

$$|u| \le \max|f| + \sqrt{2\phi}\,\{\max|g| + \max|\bar{g}|\} . \tag{4.5}$$

So we wish to find $\max|\bar{g}|$.

Although f' will make random jumps at the corners, f will be continuous at each corner, since u is to be continuous up to and including the perimeter.

In the sequel, we will use superscripts T, B, L, and R to signify the top, bottom, left, and right sides of the rectangle. We will express $\bar{u}$ as

(4.6) $$\bar{u} = \sum + \sum^T + \sum^B + \sum^L + \sum^R .$$

In this we choose $\sum$ a polynomial

(4.7) $$\sum = A + Bx + Cy + Dxy ,$$

with A, B, C, and D chosen so that $\sum$ has the same value as $\bar{u}$ at each of the four corners of the rectangle. To accomplish this, we set

(4.8) $$A = f(0,0) ,$$

(4.9) $$B = \frac{f(a,0) - f(0,0)}{a} ,$$

(4.10) $$C = \frac{f(0,b) - f(0,0)}{b} ,$$

(4.11) $$D = \frac{f(a,b) - f(a,0) - f(0,b) + f(0,0)}{ab} ;$$

recall that $f(x,y)$ is the value assumed by $\bar{u}$ around the perimeter of the rectangle.

We will first determine the inward normal for $\bar{u}$ along the top of the rectangle. A similar analysis will apply to each of the other three sides.

We have

(4.12) $$-\frac{\partial}{\partial y}\sum = -C - Dx .$$

At $x = 0$, this is

(4.13) $$-\frac{f(0,b) - f(0,0)}{b}$$

and at $x = a$, this is

(4.14) $$-\frac{f(a,b) - f(a,0)}{b} .$$

As the right side of (4.12) is linear, its extreme values must be (4.13) and (4.14). However, (4.13) is bounded below and above by the minimum and maximum of f', evaluated for $x = 0$. (Here we are considering arc length as increasing counterclockwise around the rectangle, so that f' evaluated for $x = 0$ is

$$-\frac{\partial}{\partial y} f(0,y) \text{ .)}$$

Similarly, (4.14) is bounded below and above by the minimum and maximum of $-f'$, evaluated for $x = a$.

We take

$$F = \max|f| , \tag{4.15}$$

$$F' = \max|f'| , \tag{4.16}$$

$$F'' = \max|f''| , \tag{4.17}$$

taken around the perimeter of the rectangle. Then we have just shown that

$$\left| - \frac{\partial}{\partial y} \Sigma \right| \leq F' . \tag{4.18}$$

We take

$$\Sigma^T = \sum_{j=1}^{\infty} A_j^T \frac{\sinh \frac{j\pi y}{a}}{\sinh \frac{j\pi b}{a}} \sin \frac{j\pi x}{a} , \tag{4.19}$$

$$\Sigma^B = \sum_{j=1}^{\infty} A_j^B \frac{\sinh \frac{j\pi(b - y)}{a}}{\sinh \frac{j\pi b}{a}} \sin \frac{j\pi x}{a} , \tag{4.20}$$

$$\Sigma^L = \sum_{j=1}^{\infty} A_j^L \frac{\sinh \frac{j\pi(a - x)}{b}}{\sinh \frac{j\pi a}{b}} \sin \frac{j\pi y}{b} , \tag{4.21}$$

$$\Sigma^R = \sum_{j=1}^{\infty} A_j^R \frac{\sinh \frac{j\pi x}{b}}{\sinh \frac{j\pi a}{b}} \sin \frac{j\pi y}{b} , \tag{4.22}$$

where

$$A_j^T = \frac{2}{a} \int_0^a \sin \frac{j\pi x}{a} \left\{ f(x,b) - f(0,b) - x \frac{f(a,b) - f(0,b)}{a} \right\} dx , \tag{4.23}$$

$$A_j^B = \frac{2}{a} \int_0^a \sin \frac{j\pi x}{a} \left\{ f(x,0) - f(0,0) - x \frac{f(a,0) - f(0,0)}{a} \right\} dx , \tag{4.24}$$

$$A_j^L = \frac{2}{b} \int_0^b \sin \frac{j\pi y}{b} \left\{ f(0,y) - f(0,0) - y \frac{f(0,b) - f(0,0)}{b} \right\} dy , \tag{4.25}$$

$$A_j^R = \frac{2}{b} \int_0^b \sin \frac{j\pi y}{b} \left\{ f(a,y) - f(a,0) - y \frac{f(a,b) - f(a,0)}{b} \right\} dy . \tag{4.26}$$

We note that $\sum^T$ is zero on the sides and bottom of the rectangle, and on the top it equals

$$f(x,b) - f(0,b) - x\,\frac{f(a,b) - f(0,b)}{a} ,$$

which is the value of

$$\bar{u} - \sum$$

along the top of the rectangle. Similar considerations apply to $\sum^B$, $\sum^L$, and $\sum^R$ with respect to other sides of the rectangle. So

$$\sum^T + \sum^B + \sum^L + \sum^R$$

takes the same values as

$$\bar{u} - \sum$$

around the perimeter of the rectangle. So

$$\sum + \sum^T + \sum^B + \sum^L + \sum^R \tag{4.27}$$

takes the same values as $\bar{u}$ around the perimeter of the rectangle. But (4.27) is harmonic. By the uniqueness of the solution of a harmonic equation, we conclude that (4.6) holds.

We can integrate by parts in (4.23), (4.24), (4.25), and (4.26) to get

$$A_j^T = -\frac{2a}{j^2\pi^2}\int_0^a f''(x,b)\sin\frac{j\pi x}{a}\,dx , \tag{4.28}$$

$$A_j^B = -\frac{2a}{j^2\pi^2}\int_0^a f''(x,0)\sin\frac{j\pi x}{a}\,dx , \tag{4.29}$$

$$A_j^L = -\frac{2b}{j^2\pi^2}\int_0^b f''(0,y)\sin\frac{j\pi y}{b}\,dy , \tag{4.30}$$

$$A_j^R = -\frac{2b}{j^2\pi^2}\int_0^b f''(a,y)\sin\frac{j\pi y}{b}\,dy , \tag{4.31}$$

where in (4.28) and (4.29) the double primes indicate the second partial derivatives with respect to x, and in (4.30) and (4.31) the double primes indicate the second partial derivatives with respect to y. Be it recalled that these second derivatives were assumed to be of bounded variation. By the result on p. 172 of Whittaker and Watson [3], the integrals on the right sides of (4.28), (4.29), (4.30), and (4.31) each decrease of the order of j^{-1}. So each of $|A_j^T|$, $|A_j^B|$, $|A_j^L|$, and $|A_j^R|$ goes to zero of the order of j^{-3} as j goes to infinity.

Thus the various series on the right of (4.19), (4.20), (4.21), and (4.22) converge absolutely and uniformly everywhere in the rectangle, <u>including the perimeter</u>, since

$$\left| \frac{\sinh \frac{j\pi y}{a}}{\sinh \frac{j\pi b}{a}} \sin \frac{j\pi x}{a} \right| \le 1 ,$$

etc. This assures the continuity of $\sum^T$, $\sum^B$, $\sum^L$, and $\sum^R$. If we take $\partial/\partial x$ or $\partial/\partial y$ of $\sum^T$, $\sum^B$, $\sum^L$, or $\sum^R$, we will multiply terms by a constant times j, and replace some sines by cosines, or some sinh's by cosh's; the latter will not make an appreciable difference for large j. After multiplication by j, the coefficients will still go to zero of the order of j^{-2}. This will still assure absolute and uniform convergence everywhere in the rectangle, <u>including the perimeter</u>, so that the partial derivatives will be continuous. Needless to say, $\sum$ and its first partial derivatives are continuous. So, by (4.6), $\bar{u}$ and its first partial derivatives are continuous up to and including the perimeter.

Along the top of the rectangle, the inward normal for $\sum^T$ is

$$(4.32) \qquad -\frac{\partial}{\partial y} \sum^T = -\sum_{j=1}^{\infty} \frac{j\pi}{a} A_j^T \frac{\cosh \frac{j\pi b}{a}}{\sinh \frac{j\pi b}{a}} \sin \frac{j\pi x}{a} .$$

We split the right side of (4.32) into

$$\sum\nolimits_1 + \sum\nolimits_2 ,$$

where

$$(4.33) \qquad \sum\nolimits_1 = -\sum_{j=1}^{\infty} \frac{j\pi}{a} A_j^T \sin \frac{j\pi x}{a} ,$$

$$(4.34) \qquad \sum\nolimits_2 = -\sum_{j=1}^{\infty} \frac{j\pi}{a} \frac{2A_j^T}{\exp\left(\frac{2j\pi b}{a}\right) - 1} \sin \frac{j\pi x}{a} .$$

Let $0 \le r \le 1$. Then, since $|jA_j^T|$ goes to zero of the order of j^{-2}, we conclude

$$\sum\nolimits_1 = \lim_{r \to 1} \sum\nolimits_r ,$$

where

$$\sum\nolimits_r = -\sum_{j=1}^{\infty} \frac{j\pi r^j}{a} A_j^T \sin \frac{j\pi x}{a} .$$

By (4.28), we have for $0 \leq r < 1$

$$\sum_r = \frac{2}{\pi}\int_0^a f''(t,b)\left\{\sum_{j=1}^{\infty} \frac{r^j \sin\frac{j\pi t}{a}\sin\frac{j\pi x}{a}}{j}\right\}dt .$$

This gives

$$\sum_r = \frac{1}{\pi}\int_0^a f''(t,b)\left\{\sum_{j=1}^{\infty} \frac{r^j \cos\frac{\pi j(t-x)}{a}}{j} - \sum_{j=1}^{\infty} \frac{r^j \cos\frac{\pi j(t+x)}{a}}{j}\right\}dt .$$

By (2.3), we get

$$\sum_r = \frac{1}{2\pi}\int_0^a f''(t,b)\left\{\ell n\left(1 - 2r\cos\frac{\pi(t+x)}{a} + r^2\right)\right.$$

$$\left. - \ell n\left(1 - 2r\cos\frac{\pi(t-x)}{a} + r^2\right)\right\}dt .$$

As $f''(t,b)$ is of bounded variation, we easily justify taking the limit as $r \to 1$. This gives

$$\sum_1 = \frac{1}{2\pi}\int_0^a f''(t,b)\left\{\ell n\left(1 - \cos\frac{\pi(t+x)}{a}\right) - \ell n\left(1 - \cos\frac{\pi(t-x)}{a}\right)\right\}dt . \tag{4.35}$$

We have

$$\left|\frac{1}{2\pi}\int_0^a f''(t,b)\,\ell n\left(1 - \cos\frac{\pi(t+x)}{a}\right)dt\right|$$

$$= \left|\frac{a}{2\pi^2}\int_{\pi x/a}^{\pi+(\pi x/a)} f''\left(\frac{as}{\pi} - x, b\right)\ell n(1 - \cos s)ds\right|$$

$$\leq \frac{F''a}{2\pi^2}\int_{\pi x/a}^{\pi+(\pi x/a)} \left|\ell n(1 - \cos s)\right|ds .$$

We have also

$$\left|\frac{-1}{2\pi}\int_0^a f''(t,b)\,\ell n\left(1 - \cos\frac{\pi(t - x)}{a}\right)dt\right|$$

$$= \left|\frac{-a}{2\pi^2}\int_{-\pi x/a}^{\pi-(\pi x/a)} f''\left(\frac{as}{\pi} + x,b\right)\ell n(1 - \cos s)ds\right|$$

$$\leq \frac{F''a}{2\pi^2}\int_{-\pi x/a}^{\pi-(\pi x/a)} |\ell n(1 - \cos s)|ds$$

$$= \frac{F''a}{2\pi^2}\int_{\pi+(\pi x/a)}^{2\pi+(\pi x/a)} |\ell n(1 - \cos s)|ds .$$

Adding these gives

$$|\textstyle\sum_1| \leq \frac{F''a}{2\pi^2}\int_0^{2\pi} |\ell n(1 - \cos s)|ds .$$

By (2.6) and (2.9), we have

$$|\textstyle\sum_1| \leq \frac{4CF''a}{\pi^2} . \qquad (4.36)$$

We turn to $\sum_2$. We have, of course,

$$0 \leq \frac{\frac{2j\pi b}{a}}{\exp\left(\frac{2j\pi b}{a}\right) - 1} \leq 1 .$$

So, by (4.34)

$$|\textstyle\sum_2| \leq \frac{1}{b}\sum_{j=1}^{\infty} |A_j^T| .$$

Then, by (4.28),

$$|\textstyle\sum_2| \leq \frac{2a}{\pi^2 b}\sum_{j=1}^{\infty}\frac{1}{j^2}\int_0^a |f''(x,b)|dx .$$

So

$$|\textstyle\sum_2| \leq \frac{a^2F''}{3b} .$$

Combining with (4.36) gives

$$(4.37) \qquad \left| -\frac{\partial}{\partial y} \sum^{T} \right| \le \left\{ \frac{4Ca}{\pi^2} + \frac{a^2}{3b} \right\} F'' .$$

Along the top of the rectangle, the inward normal derivative for $\sum^B$ is

$$(4.38) \qquad -\frac{\partial}{\partial y} \sum^{B} = \sum_{j=1}^{\infty} \frac{j\pi}{a} \frac{A_j^B}{\sinh \frac{j\pi b}{a}} \sin \frac{j\pi x}{a} .$$

We have, of course,

$$0 \le \frac{\frac{j\pi b}{a}}{\sinh \frac{j\pi b}{a}} \le 1 .$$

So

$$\left| -\frac{\partial}{\partial y} \sum^{B} \right| \le \frac{1}{b} \sum_{j=1}^{\infty} |A_j^B| .$$

Then, by (4.29)

$$\left| -\frac{\partial}{\partial y} \sum^{B} \right| \le \frac{2a}{\pi^2 b} \sum_{j=1}^{\infty} \frac{1}{j^2} \int_0^a |f''(x,0)| dx .$$

So

$$(4.39) \qquad \left| -\frac{\partial}{\partial y} \sum^{B} \right| \le \frac{a^2}{3b} F'' .$$

Along the top of the rectangle, the inward normal derivative for $\sum^L$ is

$$(4.40) \qquad -\frac{\partial}{\partial y} \sum^{L} = -\sum_{j=1}^{\infty} \frac{j\pi}{b} A_j^L \frac{\sinh \frac{j\pi(a-x)}{b}}{\sinh \frac{j\pi a}{b}} (-1)^j .$$

By (4.30), this gives for $0 < x \le a$

$$(4.41) \qquad -\frac{\partial}{\partial y} \sum^{L} = \frac{2}{\pi} \int_0^b f''(0,y) \left\{ \sum_{j=1}^{\infty} \frac{\sinh \frac{j\pi(a-x)}{b}}{\sinh \frac{j\pi a}{b}} \frac{(-1)^j \sin \frac{j\pi y}{b}}{j} \right\} dy .$$

Temporarily denote the material in the curly brackets by $\sum^*$. By (2.11), it is just the value at (x,y) of the harmonic function in the rectangle which equals $-\pi y/2b$ on the left side and is zero on the other three sides. So, by the principle of the maximum, $\sum^*$ is nonpositive. However

$$\sum^* + \frac{\pi y(a-x)}{2ab}$$

is a harmonic function which is nonnegative on the perimeter. So by the principle of the minimum,

$$\sum^* + \frac{\pi y(a-x)}{2ab}$$

is nonnegative. So we conclude that

$$-\frac{\pi y(a-x)}{2ab} \leq \sum^* \leq 0 .$$

Using this in (4.41) gives

$$(4.42) \qquad \left| -\frac{\partial}{\partial y} \sum^L \right| \leq \frac{a-x}{ab} \int_0^b |f''(0,y)| y dy \leq \frac{b(a-x)F''}{2a} .$$

Our proof of this required the assumption $0 < x \leq a$ to insure convergence rapid enough to permit interchange of the order of summation and integration in going from (4.40) to (4.41). But

$$\frac{\partial}{\partial y} \sum^L$$

is continuous for $0 \leq x \leq a$. So (4.42) must hold also for $x = 0$.

A similar argument will give

$$(4.43) \qquad \left| -\frac{\partial}{\partial y} \sum^R \right| \leq \frac{bxF''}{2a} .$$

So along the top of the rectangle, we have the inward normal derivative bounded as follows

$$(4.44) \qquad \left| -\frac{\partial}{\partial y} \bar{u} \right| \leq F' + \left\{ \frac{4Ca}{\pi^2} + \frac{2a^2}{3b} + \frac{b}{2} \right\} F'' .$$

Along the bottom, one gets the same bound. On each side, one gets the same bound, except with a and b interchanged.

So, for a bound on our normal derivative, we have to use whichever is larger of

$$(4.45) \qquad \frac{4Ca}{\pi^2} + \frac{2a^2}{3b} + \frac{b}{2}$$

and

$$(4.46) \qquad \frac{4Cb}{\pi^2} + \frac{2b^2}{3a} + \frac{a}{2} .$$

If we subtract (4.46) from (4.45), the difference is seen to be

$$\frac{(a-b)}{6ab}\left\{4a^2 + ab + 4b^2 + \frac{24Cab}{\pi^2}\right\} .$$

Hence, we see that if $a \geq b$, then (4.45) is greater than or equal to (4.46). So we get the following result:

Theorem 2. Let u be a biharmonic function satisfying (1.1), (1.2), and (1.3) in a rectangle having sides of lengths a and b, where $a \geq b$. Let u and its first partial derivatives be continuous up to and including the perimeter. On each side, let f'' be of bounded variation in the closed interval consisting of that side. At each corner, let there be a neighborhood of the corner within which $\nabla^2 u$ is bounded. Then

$$(4.47) \qquad |u| \leq \max|f| + \sqrt{2\phi}\left\{\max|g| + \max|f'| + \left[\frac{4Ca}{\pi^2} + \frac{2a^2}{3b} + \frac{b}{2}\right]\max|f''|\right\} ,$$

where the maxima are taken over the perimeter, and $\phi(x,y)$ is the function which satisfies

$$(4.48) \qquad \nabla^2\phi = -1$$

inside the rectangle and has $\phi \equiv 0$ on the perimeter, and C is given by (2.7).

NOTE. If the rectangle is oriented with one corner at the origin, one side of length a along the positive x-axis, and another side of length b along the positive y-axis, then ϕ will be defined as in the paragraph beginning just before (4.4). So ϕ is nonnegative and is bounded by

$$\frac{1}{2}\max\{x(a-x),\ y(b-y)\} .$$

5. POSSIBLE WEAKENING OF HYPOTHESES. In Thm. 2 we assume continuity of both first derivatives up to and including the perimeter. Could we relax this assumption just at the four corners?

To get some feeling for this, consider the unit square, ABCD (see Fig. 1), situated in the first quadrant with A at the origin. Let us require that u be zero on the perimeter, and ask for an inward normal of -1 along AB and DC and of +1 along AD and BC.

Lift the figure out of the plane, and flip it about the diagonal AC; AB goes up and over to the positive of AD, while AD goes down and under to the position of AB. We now have the same boundary conditions that we started with. So, if they determine u uniquely, we must have the same values of u

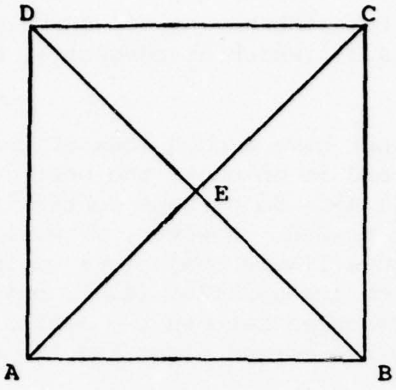


Figure 1

as before. However, since the figure has been flipped upside down, the values along AC have been changed to their negatives. As they come out the same as before, they must be zero. Similarly, we conclude that u is zero along BD.

As we have an inward normal of -1 along AB, the values of u next to AB must be negative. Likely u is negative all inside the triangle AEB, sloping down from AB and up to AE and EB. However, even if this is not the case, consider what happens if one starts vertically from AB, at a distance x from A with $x < 1/2$. One starts off with a slope of -1, which certainly takes one to negative values of u. But by the time one gets up to AE, u has got back up to zero. So one must encounter some place of positive slope. So the slope has gone from -1 to a positive value in a distance less than x. So, someplace along the way $\partial^2 u/\partial y^2$ must be at least $1/x$.

To find out what $\nabla^2 u$ is doing, we have also to get an idea of the behavior of $\partial^2 u/\partial x^2$. Let us go along parallel to AB, and close to it, from AE to BE. We start with $u = 0$ at AE and finish with $u = 0$ at BE. If we are close enough to AB, u will be mostly negative in between. This indicates that $\partial^2 u/\partial x^2$ will tend to be positive.

Thus it appears that as we approach A in the triangle AEB, we will encounter points where $\nabla^2 u$ is greater than $1/x$. Not only are we violating our Boundedness Hypothesis, but it appears possible that $u\nabla^2 u$ is not approaching zero as we get close to A. So Miranda's Lemma is likely failing. This voids our proof of Theorem 2.

This suggests that if we admit discontinuity of first derivatives at a corner, we may entail a violation of our Boundedness Criterion.

If we retain continuity of both first derivatives at the corners, do we really need the Boundedness Hypothesis? The key result is (4.3). As $u - \bar{u}$ is zero along the perimeter, the derivative along the perimeter must also be zero. Given continuity at the corner, $u - \bar{u}$ and both its first derivatives must approach zero continuously as one approaches a corner. This does not seem to leave much latitude for misbehavior of $u - \bar{u}$. We will conjecture that this suffices to give (4.3) (which is adequate), but do not see at this time how to prove it.

Actually, though we do not have a good idea of the behavior of the u of Fig. 1 (assuming it exists, and is unique), the best guess we can make indicates that it actually satisfies (1.4). So perhaps continuity of first derivatives at the corners is not really needed. However, we will not venture to conjecture this. In view of (3.5), a more likely conjecture would be that one could prove something like (1.4), or its parallel (4.3), but with an extra term on the right involving the differences between the limits at a corner of a first derivative as one approaches the corner along the two sides.

REFERENCES

[1] M. Abramowitz and I. A. Stegun, "Handbook of Mathematical Functions," National Bureau of Standards, Applied Mathematics Series No. 55, U. S. Government Printing Office, Washington, D. C. 1964.

[2] C. Miranda, "Formule di maggiorazione e teorema di esistenza per le funzioni biarmoniche di due variabili," Giornale di Matematiche Battaglini, vol. 78 (1948-49), pp. 97-118.

[3] E. T. Whittaker and G. N. Watson, "A Course of Modern Analysis," 4-th edition, Cambridge University Press, 1952.

[4] G. Lauricella, "Sull'integrazione dell'equazione $\Delta\Delta u = 0$ in un campo circolare," Atti. della R. Accademia delle Scienze di Torino, vol. 31 (1895-96), p. 1010.

[5] M. Picone, "Nuovi indirizzi di ricerca nella teoria e nel calcolo delle soluzioni di talune equazioni lineari alle derivate parziali della Fisica-matematica," Annali della Scuola Norm. Superiore de Pisa, series 2^a, vol. 5 (1936), pp. 213-288.

A NUMERICAL METHOD FOR LARGE STIFF SYSTEMS OF ORDINARY DIFFERENTIAL EQUATIONS

T. P. Coffee, J. M. Heimerl, and M. D. Kregel
Ballistic Modeling Division
US Army Ballistic Research Laboratory
Aberdeen Proving Ground, MD 21005

ABSTRACT. A method is described for the efficient integration of large stiff systems of ordinary differential equations. The method is based on a predictor-corrector formulation, that uses a very accurate predictor and evaluates the Jacobian in a non-standard fashion. The resulting program is compared with EPISODE, a standard integrator based on Gear's stiff formulas, for a number of systems of ordinary differential equations. The results show that the procedure is competitive with EPISODE, and is much more efficient for some problems.

I. INTRODUCTION.

We will consider the problem of solving a set of stiff ordinary differential equations of the form

$$y_k' = F_k(y_1, y_2, \ldots y_m) \quad k=1,2,\ldots,m \tag{1}$$

$$y_k(t_o) = y_{ko}$$

By stiff, we mean that the equations have widely different time constants, several orders of magnitude apart.

Standard numerical methods are constrained to a step size roughly comparable to the smallest time constant. As a result, a prohibitive number of integration steps will be required to solve the system. So different methods have been developed for stiff systems.

A common approach is the predictor-corrector formulation. We attempt to predict y_k at the n'th time step using an explicit formula of the form

$$y_{kn}^P = \sum_{i=1}^{P} \alpha_i\, y_{k(n-i)} + h \sum_{i=1}^{P} \beta_i y_{k(n-i)}' . \tag{2}$$

The predictor can be made as accurate as necessary, but will not have sufficient stability. That is, when a reasonable step size is used, errors will grow and eventually become larger than the real solution.

So an implicit corrector is used. We compute

$$y_{kn}^{P'} = F_k(y_{1n}^P, y_{2n}^P, \ldots, y_{mn}^P). \tag{3}$$

Then

$$y^{C}_{kn} = \sum_{i=1}^{P} \alpha_i y_{k(n-i)} + h \sum_{i=1}^{P} \beta_i y'_{k(n-i)} + h\beta_o y^{P'}_{kn} \quad . \tag{4}$$

The corrector formula is chosen to have adequate stability.

We want to find a final value y^{F}_{kn} such that

$$y^{F}_{kn} = \sum_{i=1}^{P} \alpha_i y_{k(n-i)} + h \sum_{i=1}^{P} \beta_i y'_{k(n-i)} + h\beta_o y^{F'}_{kn} \ . \tag{5}$$

Simply iterating the corrector will not converge, unless the time step is very small. But we can approximate $y^{F}_{kn} - y^{C}_{kn}$ by a truncated series expansion. Using the notation

$$y^{F}_{kn} = y^{P}_{kn} + d_k \ , \tag{6}$$

this results in the equations

$$y^{P}_{kn} - y^{C}_{kn} = h\beta_o \sum_{j=1}^{m} \left(\frac{\partial y^{P'}_{kn}}{\partial y^{P}_{jn}} - \frac{\delta_{kj}}{h\beta_o}\right) d_j . \tag{7}$$

So we have a set of m linear equations in the m unknowns d_k, involving the y^{P}_{kn}, the y^{C}_{kn}, and the Jacobian elements $\frac{\partial y^{P'}_{kn}}{\partial y^{P}_{jn}}$. This process is called Newton-Raphson iteration.

Solving such a system involves approximately $m^3/3$ multiplications and divisions. For larger systems of equations, most of the computation time is involved in solving these equations. So a basic problem in solving large systems is obtaining an efficient procedure for handling these equations and the corresponding matrix of Jacobian elements.

II. THE METHOD OF GEAR. The algorithm DIFSUB, by C. W. Gear (1), marks an important advance in numerically solving stiff systems. The program actually consists of two integrators, one for stiff problems and one for non-stiff problems.

The stiff integrator uses predictors of the form

$$y^{P}_{kn} = \sum_{i=1}^{P} \alpha_i y_{k(n-i)} + h\beta_1 y'_{k(n-1)} \qquad P = 1,2,\ldots, 5 \tag{8}$$

and correctors of the form

$$y_{kn}^{C} = \sum_{i=1}^{P} \alpha_i y_{k(n-i)} + h\beta_o y_{kn}^{P'} \qquad P = 1,2,\ldots, 5. \tag{9}$$

The higher order formulas are more accurate, while the lower order formulas are more stable. The algorithm changes the order automatically during the integration.

DIFSUB does not reevaluate the Jacobian every step. It first attempts to find the correction factors d_k using the previous Jacobian values. If it fails to obtain convergence in three Newton-Raphson iterations, the Jacobian is reevaluated. This can occur as often as every step or only every three or four steps, depending on how rapidly the Jacobian changes. For larger systems, this can result in a substantial savings in computation time.

A number of variants of DIFSUB are now in existence. Most of these are available from Argonne Laboratories upon request.

In particular, one version (EPISODE) also varies the step size dynamically. Most predictor-corrector formulas use fixed α's and β's, and assume that all previous steps were of the same size. To change the step size, it is normally doubled or halved, using interpolation if necessary to find the appropriate values. In EPISODE, α's and β's can be modified, and any succession of step sizes can be used. This is especially important in integrating past discontinuities, where the step size must be drastically reduced. This problem will be discussed below.

III. THE K-INTEGRATOR. The K-integrator, developed at the Ballistic Research Laboratory, is also a predictor-corrector method. Its corrector is a fixed, third order, three step formula

$$y_{kn}^{C} = \sum_{i=1}^{3} \alpha_i y_{k(n-i)} + \sum_{i=1}^{2} \beta_i y'_{k(n-i)} + h\beta_o y_{kn}^{P'} . \tag{10}$$

The exact formula was arrived at empirically. Its stability and accuracy characteristics are roughly comparable to Gear's third order formula.

Like EPISODE, it is a variable step size program. Unlike the Gear programs, the K-integrator does not vary the order of the corrector. Our results indicate that being able to vary the order is much less important than the method of treating the Jacobian and the step size changing algorithm.

The main difference between the K-integrator and DIFSUB is an initial screening. Before generating the Jacobian, we check the agreement of the predictor and corrector and the stiffness of each equation.

More specifically, we let $r_k = - \frac{\partial y_k'}{\partial y_k}$. r_k is the inverse of the time constant of the equation. Then if

$$|y_{kn}^C - y_{kn}^P| < \text{error bound} \tag{11}$$

and

$$|h\, r_{k(n-1)}| < 1, \tag{12}$$

we accept y_{kn}^C as y_{kn}^F.

For each element that has converged, the corresponding row and column in the Jacobian matrix can easily be eliminated. There are no stability problems, since only the non-stiff elements are eliminated. The smaller matrix is solved explicitly each step. This can result in substantial savings, depending on how many elements can be eliminated in the initial screening.

Because of this initial step, we want our predictor to be as accurate as possible. So a non-standard form for the predictor is used.

We will write equations (1) in the form

$$y_k' = f_k - r_k y_k \tag{13}$$

where f_k and r_k are generated internally by the program, using the rules

$$f_k = y_k' - \frac{\partial y_k'}{\partial y_k}\, y_k \tag{14}$$

$$r_k = - \frac{\partial y_k'}{\partial y_k}\,. \tag{15}$$

In general, f_k and r_k vary less rapidly than y_k. (Most stability analyses assume these functions are constant.)

So we first predict f_k and r_k, using the simple second order, three step formulas

$$f_{kn}^P = \sum_{i=1}^{3} \alpha_i\, f_{k(n-i)} \tag{16}$$

$$r_{kn}^P = \sum_{i=1}^{3} \alpha_i\, r_{k(n-i)}. \tag{17}$$

Then

$$y'_{kn} \approx f^P_{kn} - r^P_{kn}\, y^P_{kn} .$$

Substituting into the corrector (10) and solving for y^P_{kn} results in

$$y^P_{kn} = \frac{\sum_{i=1}^{3} \alpha_i y_{k(n-i)} + h \sum_{i=1}^{2} \beta_i y'_{k(n-i)} + h\beta_o f^P_{kn}}{1 + h\beta_o\; r^P_{kn}} \qquad (18)$$

This is our predictor.

It can be shown that this is a third order formula. Under the common assumption that f and r are constant, it has the same stability as the corrector.

IV. COMPARISON WITH EPISODE. A number of comparisons were made on a CDC 7600 in single precision between the K-integrator and EPISODE. Five small problems were chosen from a set proposed by Enright, Hull, and Lindberg (2). These problems, A4, B4, C5, D6, and E5, in their article cover a wide variety of types ranging in size from three to ten ordinary differential equations.

The average of the results is given in Table 1. While an average will obscure any problem dependent features, it does show that the methods are roughly comparable. (Details will be included in a future BRL report.)

EPISODE is more efficient at a stricter error tolerance, where more steps are taken. The Jacobian ages less rapidly, and so does not need to be reevaluated as often.

The K-integrator is less accurate at the stricter error tolerance. The program was developed to deal with chemical kinetics problems, where appropriate rate constants are known only approximately. So the step size changing algorithm was developed to handle problems requiring only moderate accuracy. At the stricter error criterion, where accuracy can be as important as stability, the K-integrator is overly optimistic.

In general, the K-integrator requires fewer steps than EPISODE. Its reduced Jacobian is somewhat more accurate than EPISODE's aged Jacobian.

The largest program run is an atmospheric model of charge flow among 64 species under the influence of a large electron flux. About 500 reactions are involved. The details are given by Heimerl and Niles (3). A preprocessor writes the subroutines for computing the derivatives and Jacobian elements (4). Of practical interest is the integration to 10^4 seconds. The results are given in Table 2.

TABLE 1

Error Criterion = 10^{-2}

	Run Time	No. of Steps	Error X 10^2
K	.015	40	.01
E	.022	49	.18

Error Criterion = 10^{-4}

	Run Time	No. of Steps	Error X 10^4
K	.029	77	.34
E	.048	100	.55

Error Criterion = 10^{-6}

	Run Time	No. of Steps	Error X 10^6
K	.076	206	4.07
E	.091	212	.60

TABLE 2

Error Criterion = 10^{-2}

	Run Time	No. of Steps
K	11.749	333
E	14.820	522

The K-integrator is somewhat faster. However, the programs actually exhibit quite different behaviors. Table 3 breaks the integration into two parts, from 0 to 10^{-4} seconds and from 10^{-4} to 10^{4} seconds, and gives the corresponding run times.

TABLE 3

	0 - 10^{-4} sec.	10^{-4} - 10^{4} sec.
K	1.6 sec.	10.1 sec.
E	8.1 sec.	6.7 sec.

At the start, the K-integrator is much faster. Only a few of the equations are stiff, and the corresponding reduced matrices are quite small.

Nearer equilibrium, most of the equations are stiff, and the K-integrator is inverting much larger matrices. However, the Jacobian changes slowly, and EPISODE's strategy of using an aged Jacobian is more efficient.

So the efficiency of the two methods for large systems is very problem dependent.

A variable step size program should be able to handle discontinuities. So the above problem was run with a square wave driving function (Figure 1). The discontinuities occur at the powers of 10.

The K-integrator reached the value of 10^{4}, while EPISODE could not get past the discontinuity at 10^{3}. A comparison of the run times is given at t = 500 seconds in Table 4.

TABLE 4

t = 500 seconds

	Run Time	No. of Steps
K	28.2	569
E	61.3	1376

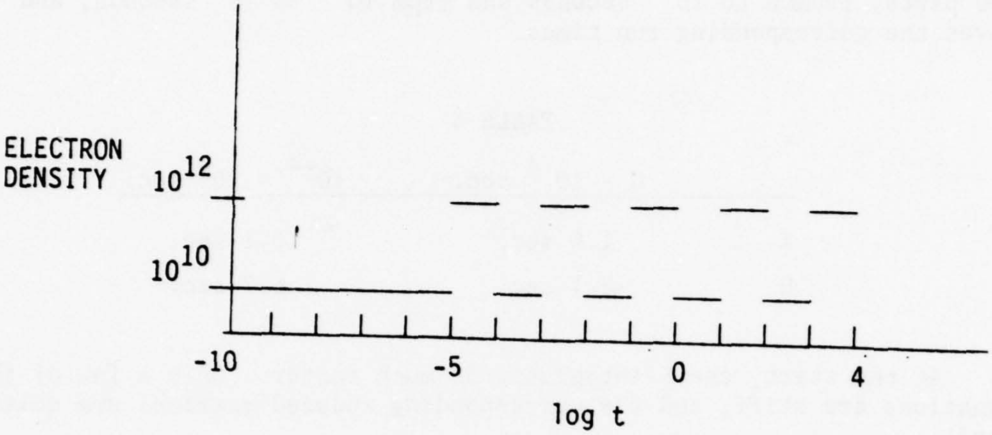


Figure 1. Square wave driving or source function; electron density (cm^{-3}) vs. log of time (seconds).

The difficulty occurs in the step size changing algorithm. EPISODE compares the original predictor with the final accepted value, then uses this value to determine the size of the next time step. But at a discontinuity, an explicit predictor is very inaccurate, and the step size is reduced drastically. At $t = 10^3$, the step size becomes 10^{-12}. But on a CDC 7600 in single precision $10^3 + 10^{-12} = 10^3$, and no further progress is possible.

The K-integrator instead compares the corrector with the final value. This is less conservative, but normally creates no problems.

So the K-integrator does seem to be comparable to the multi-order EPISODE program, at least for a looser error criterion. It can be more efficient, depending on the problem, and is especially good over discontinuities.

REFERENCES

1. C. W. Gear, C. ACM 14, (1971), 185.

2. W. H. Enright, J. C. Hall, and B. Lindberg, BIT 15 (1975), 10.

3. J. M. Heimerl and F. E. Niles, "BENCHMARK-76: Model Computations for Disturbed Atmospheric Conditions. 1. Input Parameters," BRL Report No. 2022, October 1977.

4. M. D. Kregel, J. M. Heimerl and E. L. Lortie, "LOADER: A Character Manipulation Program for Automatic Code Writing," BRL Report No. 1814, August 1975.

NONLINEAR REALIZATION THEORY

R. E. Kalman
Center for Mathematical System Theory
University of Florida
Gainesville, FL 32611

ABSTRACT. This talk is intended to give a summary of current results and problems in realization theory, with special attention to the nonlinear problem. Since the results of linear (finite-dimensional) realization theory may be regarded as already classical in 1978, current efforts are directed toward the understanding of nonlinear realization problems, and especially the development of methodology which provides nontrivial generalizations of the insights gained in the linear case. Consequently the research is heavily algebraic in character. Most of the remarks will refer to work in progress or published at the Center for Mathematical System Theory in Gainesville.

1. REALIZATION THEORY = BASIC PROBLEM IN SYSTEM THEORY. It has become conventional to base the (rigorous) mathematical definition of a dynamical system on a set of equations involving internal variables. This point of view is usually (but not necessarily always) the most efficient one also for applied mathematical work, such as optimization, stability analysis, etc. On the other hand, empirical data are usually presented in terms of an external definition of a system (impulse response, transfer function, and other input/output relations). The connecting link between these two points of view is technically known as realization theory.

Given an internal description of a system, deducing its external description is a purely mathematical (computational) problem. The converse is nontrivial: given a system presented by its external description, the construction of its internal structure is conceptually the same as scientific model building from experimental data.

Realization theory is concerned with "automating" this process; that is, showing how models can be built from data by means of a purely mechanical, deductive procedure. To carry out this program requires considerable mathematical sophistication and reveals questions, properties, and results which are nonobvious to those who view modeling as a kind of art which should be practiced naively and intuitively.

The theoretical framework which has been developed for realization theory can hardly be attacked on conceptual or scientific grounds. However, the question of its effectiveness is highly relevant; all depends on the class of systems for which the abstract principles can be sharpened to an explicit pure-mathematical theory and practical applied-mathematical computational procedures.

The task of system theory is simply to keep enlarging the class of systems for which something useful can be said about the realization problem.

2. LINEAR REALIZATION THEORY. An important fact of linear system theory today (perhaps the most important from the point of view of the conceptually but not mathematically sophisticated user) is that there is agreement on the basic definition. As long as the words

(L) finite-dimensional, finitely many inputs, finitely many outputs, discrete-time, constant (time-invariant), real, linear

are acceptable as attributes of the class considered, a small amount of deductive axiomatics will prove that such systems Σ are (uniquely) described by the equations

(1) $x_{t+1} = Fx_t + Gu_t$,

(2) $y_t = Hx_t$,

where t is an integer (time), x, u, y are real, finite-dimensional vectors (called state, input, output), and F, G, H are matrices with real, constant coefficients. Since this description of Σ is universally adopted and unambiguous, it is possible to do mathematics by abstractly taking the triple of matrices $\Sigma = (F, G, H)$ as the fundamental object.

remembering also the transformation law

(3) $(F, G, H) \mapsto (TFT^{-1}, TG, HT^{-1})$

under the action of a basis change $x \mapsto Tx$ in the state space X. [The case of continuous time is <u>also</u> explicitly handled by abstractly viewing the triple (F, G, H) subject again to (3) but now having a different physical interpretation via the differential equation

(1') $dx/dt = Fx + Gu(t)$.]

The above is a concise summary of the <u>internal definition</u> of linear systems possessing the attributes (L). [For infinite-dimensional linear systems, the question of the appropriate definition is by no means completely settled at the present time.]

The <u>external definition</u> is based on computing the relation between the time series $\{u_t\}$ and $\{y_t\}$ for some fixed initial state, say, $x_o = 0$. An elementary calculation shows that

(4) $y_t = \sum_{t>\tau\geq 0} A_{t-\tau} u_\tau, \quad x_o = 0, \quad t = 1, 2, \ldots$,

where

(5) $A_t := HF^{t-1}G, \quad t = 1, 2, \ldots$.

If the system Σ is given via the internal equations (1-2), the evaluation of the right-hand side of (5) is all that is needed to determine the external description.

On the other hand, if Σ is given via its external description, i.e., the sequence

(6) $S = \{A_1, A_2, \ldots\}$,

then the realization problem amounts to finding matrices F, G, H which satisfy the identities (5) for all positive integer values of t. For this reason, (5) is known as the realization condition.

Since (5) involves infinitely many conditions, the theory of the realization problem for our class (L) is nontrivial; in fact, so much so, that the theory of relations (5) is actually more difficult in the finite case than in the infinite case (KALMAN [1971]).

The format of condition (5) shows immediately that the realization problem cannot have a unique solution: for example, given any triple (F, G, H) satisfying (5) for a given sequence S, we may enlarge each of these matrices arbitrarily without violating the realization condition by introducing zeroes as new entries. [Mathematical intuition suggests that padding a given realization by zeroes is not essential; other numbers could also be used, with some precautions.]

To evolve a viable theory, it is necessary to attach a realization (F, G, H) to the data S in some natural way. By a structural analysis of the system Σ we see that we can only hope to do so if the realization has the properties of "completely reachable (completely controllable) and completely observable"; this was shown in the first rigorous paper in realization theory (KALMAN [1962]). It follows, furthermore, that any realization Σ can be "reduced" to such (smaller) system $\hat{\Sigma}$ without affecting its input/output properties, i.e., the fact that $\hat{\Sigma}$ is a realization. The Uniqueness Theorem for Canonical Realizations (found in 1962, see [KALMAN, FALB, and ARBIB, 1969, Chapter 10, Appendix 10.c]) then states that all realizations which are simultaneously completely reachable and completely observable are a single isomorphism class---that is, this condition guarantees that the realization is naturally associated with the data.

An algebraic analysis of the preceding results shows that

completely reachable ≈ surjection
completely observable ≈ injection

and that the realization problem is abstractly equivalent to the canonical factorization of a vector-space (or $R[z]$-module) homomorphism. Consequently we are justified in using the technical algebraic term "canonical" for such realizations which uniquely correspond to the data.

A further analysis of the abstract algebraic setting shows that the results concerning canonical realizations are valid in the most abstract category-theoretical setting. See KALMAN [1976]. Consequently the problem of realization reduces to a technical mathematical problem, requiring, for example, the discovery of appropriate explicit conditions corresponding to "canonical".

3. UNIQUENESS THEOREM FOR CANONICAL REALIZATIONS OF POLYNOMIAL SYSTEMS.

The first nonlinear case for which the uniqueness theorem holds under an explicitly known condition for "canonical" was discovered by SONTAG and ROUCHALEAU [1976]. They consider the class of systems characterized by

(P) finite-dimensional, finitely many inputs, finitely many outputs, discrete-time, constant (time-invariant), real, polynomial

which is the same as the class (L) except that "linear" has been replaced by "polynomial". Evidently this is in some sense the simplest class of algebraically definable nonlinear systems. In terms of equations, the internal description of such a system is given by

$$x_{t+1} = f(x_t, u_t), \tag{6}$$

$$y_t = h(x_t). \tag{7}$$

Here f and h are polynomials; otherwise the system interpretation of these equations is the same as for (1-2).

The appropriate definition of "canonical" for which the uniqueness theorem holds is then

(8) canonical := quasi-reachable + algebraically observable.

By "quasi-reachable" we mean that the (Zariski) closure of all reachable states is the whole state space X. By "algebraically observable" we mean that the initial state can be recovered from finitely many output values by substituting these values into an explicit formula (evaluating a polynomial).

The theorem of SONTAG and ROUCHALEAU, while very important as a fundamental (first) result in nonlinear realization theory, is basically not a new result of pure mathematics. Rather, it is a sophisticated translation of the same theorem specialized to the linear case over to a nonlinear problem. This translation is achieved by defining certain abstract dual systems, using the classical algebraic-geometric idea of studying a variety (algebraic set) via the polynomial functions (coordinate ring) defined on it. Such a procedure yields an (abstract, nonphysical) linear system, for which the definition of "canonical" and the uniqueness theorem are classical; by backtranslating these notions via duality to the original nonlinear setting leads to definition (8) as well as to the uniqueness theorem for polynomial systems.

No cases are known at present where a uniqueness theorem for canonical realizations has been established without explicit reference to the abstract technique which proved the corresponding result in the classical linear case.

The setup used by SONTAG and ROUCHALEAU may be imitated, for category-theoretic reasons, in the infinite-dimensional linear case. See the dissertation YAMAMOTO [1978].

4. CONCRETE REALIZATION THEORY (LINEAR CASE). The main pure-mathematical problem of linear realization theory is essentially to give an explicit finiteness condition for the existence of a finite-dimensional realization; the main applied-mathematical problem is to provide an algorithm for the construction of a canonical realization.

Both problems are simultaneously solved in the linear case by the celebrated Hankel construction. Define the behavior (or Hankel) matrix B_S corresponding to the external description S (of Σ) by

$$(9) \qquad B_S := \begin{bmatrix} A_1 & A_2 & A_3 & \cdots \\ A_2 & A_3 & A_4 & \cdots \\ \cdot & \cdot & \cdot & \\ \cdot & \cdot & \cdot & \\ \cdot & \cdot & \cdot & \end{bmatrix}$$

Then we have the theorem:

$$(10) \qquad \dim \Sigma^{can}(S) = \operatorname{rank} B_S.$$

As a nontrivial consequence of this basic result, the applied mathematical problem of actually calculating the rank of B_S contains in it some partial results which, with suitable editing and data rearrangement, practically amount to the computation of the canonical realization. Thus, at least in the linear case, there is an intimate connection between determining if a finite-dimensional realization exists at all and then finding the canonical realization.

The proof of (10) is abstractly exceedingly simple. We define the state space X_S of the desired (canonical) realization of S by setting

$$(11) \qquad X_S := \text{vector space spanned by the (infinite) columns of } B_S.$$

This definition is useful only if rank B_S is finite; it has then the peculiarity that X_S, a finite-dimensional space, is generated by vectors taken from an infinite dimensional space.

This definition is intrinsic; it depends only on B_S, hence on S; that is, X_S is abstractly a function of the data S. That this state space actually works is proved by explicitly constructing (F_S, G_S, H_S) from B_S and then showing that these matrices satisfy the realization condition (5). That this can be done is a consequence of the "Hankel geometry" of B_S. In other words, the main idea of the proof is that the data S has much nicer properties if it is arranged in the Hankel pattern of an (infinite) rectangular matrix than viewed merely as an infinite sequence $(A_1, A_2, \ldots)$.

It is not well known, but easily seen, that the entire construction of a canonical realization just described "dualizes" when we replace (11) by

(11') X_S := vector space spanned by the (infinite) rows of B_S.

This second construction may be generalized more easily in the non-linear case.

5. BILINEAR RESPONSE FUNCTIONS. The actual construction of a realization in the nonlinear case depends on the ability to imitate the Hankel construction, which works so easily in the linear case.

We owe to FLIESS [1974] the remarkably surprising discovery that the concept of the Hankel matrix is totally independent from linearity and can be applied to power series in finitely many (even noncommutative) variables. This is due to the fact that the basic definition due to Hankel that

(12) $(B_S)_{\mu,\nu} := f(\mu o \nu)$

can be made whenever the abstract quantities μ, ν form a (multiplicative, but not necessarily commutative) semigroup; in the system-theoretic context this works because the right-hand side of (12) is the value of the response map of the system under the inputs μ and ν and these inputs form the desired semigroup under the usual operation of concatenation.

This so called "concatenative" definition of the generalized Hankel matrix was given a thorough theoretical study in the dissertation of SONTAG [1976]. It turns out that the rank of the (generalized) Hankel matrix is still a decisive finiteness condition. It is seen also, at least in some examples, that the matrix so defined is a very inefficient method of displaying the information relevant to realization theory, since the rank is already determined by a very small (but information-carrying) submatrix.

What the most economical and pure mathematically sharpest definition of the behavior matrix is must be considered an unsolved problem at the present time; at any rate, there are many alternatives.

6. BILINEAR RESPONSE FUNCTIONS. The first nonlinear class for which a complete realization theory is available is that given by bilinear response functions. See KALMAN [1979]. This is an external description of the system; the observed output values are bilinear functions of the inputs in two distinct channels.

For this class it is possible to define a behavior matrix which uses the convolution multiplication (rather than concatenation) to arrive at a Hankel-like definition. (The details are too complicated to be given here.) It should be noted, however, that the rigid Hankel definition cannot be used for otherwise the behavior matrix will have infinite rank; the correct behavior matrix replaces the elements defined by the Hankel rule in certain parts of the matrix simply be zeroes. In this semi-arbitrary way, the fundamental property:

finite-rank behavior matrix ≈ finite realizability

is preserved but at the expense of an ad-hoc definition of the behavior matrix. A deeper understanding of the questions is evidently lacking at the present time.

7. PEARLMAN's THEOREM. While the question of the sharpest possible generalization of the behavior matrix to the nonlinear case is still shrouded in confusion, a very interesting algebraic result was discovered in the dissertation of PEARLMAN [1976].

This result is concerned with giving an explicit algebraic criterion for quasi-reachability (necessary and sufficient). The corresponding criterion in the linear case was the result from which all of (linear) realization theory has developed, which certainly suggests that PEARLMAN's criterion should be of basic interest for nonlinear realization theory. The criterion is restricted to systems relevant to the realization of bilinear response functions; what is very surprising, however, is that the criterion is expressed via the classical reachability rank condition for linear systems.

The statement of the theorem (for complete definitions, please refer to PEARLMAN [1976]) is the following:

Let Σ *be a nonlinear system realizing a bilinear response function. Let* L_Σ *be a linear system constructed from the information defining* Σ. *Then*

Σ = *quasi-reachable iff* L_Σ = *reachable*.

The phrase "constructed from information defining Σ" means that the nonlinear system Σ is defined via a collection of matrices (just like any linear system), these matrices are then algebraically combined (via tensor products, etc.) to form the two matrices F_L and G_L required for substitution into the linear reachability criterion.

The necessity of the criterion is trivial, but the proof of sufficiency requires very lengthy direct constructions.

The theorem is remarkable in that a nonlinear problem is abstractly reduced to the solution of a linear problem, without this reduction in any way corresponding to classical mathematical ideas such as local linearization.

8. REFERENCES

M. FLIESS

[1974] "Matrices de Hankel", J. Math. Pures et Appl., 53: 197-224.

R. E. KALMAN

[1962] "Canonical structure of linear dynamical systems", Proc. Nat. Acad. Sci. (USA), 48: 596-600.

[1971] "On minimal partial realizations of a linear input/output map", in Aspects of Network and System Theory, (edited by R. E. Kalman and N. DeClaris), Holt, Rinehart, and Winston, pp. 385-408.

[1976] "Realization theory of linear dynamical systems", in Control Theory and Functional Analysis, Vol. II, International Atomic Energy Agency, Vienna, pp. 235-256.

[1979] "Realization theory of bilinear response functions", to appear.

R. E. KALMAN, P. L. FALB, and M. A. ARBIB

[1969] Topics in Mathematical System Theory, McGraw-Hill.

J. G. PEARLMAN

[1977] "Internal description of multilinear systems", Ph.D. dissertation, Imperial College.

E. D. SONTAG

[1976] "On the internal realization of polynomial response maps", doctoral dissertation, Center for Mathematical System Theory, University of Florida.

E. D. SONTAG and Y. ROUCHALEAU

[1976] "On discrete-time polynomial systems", J. Nonlinear Analysis and Applications, 1: 55-64.

Y. YAMAMOTO

[1978] "Realization theory of infinite-dimensional linear systems", doctoral dissertation, Center for Mathematical System Theory, University of Florida.

THE RADIATION PATTERN OF A DIELECTRIC ANTENNA - AN ASYMPTOTIC APPROACH

Walter Pressman
US Army Communications Research and Development Command
Fort Monmouth, New Jersey

ABSTRACT

It is possible to represent the field radiated by a dielectric antenna as a complicated one dimensional integral with highly oscillatory integrand dependent upon several parameters. By suitable transformation, dependence of the integral on one large parameter can be accomplished. Then by using asymptotic techniques the integral can be approximated. This is done for the TE Mode of operation. Although not done here, the TM-mode can be analyzed in exactly the same way. The approximations obtained are useful both for qualitative and numerical analysis of the antenna pattern.

1. INTRODUCTION

The Army is developing compact low-cost mm-wave transceivers for high data-rate communication at command posts. One approach emphasizes image and insular line technology. Dielectric antennas are directly compatible with these dielectric waveguide technologies and consequently are likely to lead to cost-effective designs. These antennas are made of non-metallic material and may take the form of a slender cone or pyramid. The theory of such structures is mathematically difficult. As a first step, an approximate theory (using a so called Local-mode approach) has been developed by the Antenna Team of Comm/ADP Laboratory for the two-dimensional problem of a wedge. The radiation pattern and directivity gain, which are extremely difficult to evaluate, lend themselves to asymptotic solution. It is proposed to determine an asymptotic method for evaluating the radiation pattern integrals in closed form; this will permit the pattern dependence on relevant engineering parameters to be established explicitly.

The relevant field integrals will be transformed so that their dependence on a large parameter becomes evident. The intrinsic characteristics of the oscillations (stationary points, coalescence, end-point behavior) will then be studied so that suitable, asymptotic techniques can be applied to determine the radiated field pattern.

The general methodology will enable the engineer to calculate the antenna pattern as a function of the important engineering variables. By varying the parameters of the analytic representation a simple method is obtained for approximating the desired antenna pattern, thereby minimizing the need for an extensive field measurement program to determine the optimum antenna characteristics.

In section two we formulate the problem mathematically and give the relationship of the pertinent parameters. In section three we perform a detailed analysis for the transverse electric (TE) mode of polarization. The critical points of the integrand are determined and then the appropriate asymptotic techniques are used to approximate the radiation pattern integral. In section four we make some brief comments on the physical behavior of the antenna and the transverse magnetic (TM) case of antenna polarization.

2. FORMULATION OF PROBLEM

We consider a dielectric antenna in the shape of a three-dimensional wedge whose cross-section in the (X Y) plane is constant with respect to Z. Thus we will consider a two-dimensional problem. Figure 1 shows the antenna which extends in the X-direction from zero to L. The transmission line feeding the antenna lies beyond L. The variable antenna cross-section is designated by $\delta(x)$ and ϕ is the angle at which energy is radiated from the antenna, measured with respect to the antenna axis. We will consider one mode of operation: TE polarization. There are four variable parameters: α, β, δ and τ whose relationships follow. The parameters α and β are are related by

$$\beta = (n^2 - \alpha^2)^{1/2} \quad , \quad \alpha, \beta \geqq 0. \qquad (2.1)$$

Here α and β are the normalized propagation constants inside the wave guide in the X and Y directions respectively, and n is the constant index of refraction

$$n^2 = 12. \qquad (2.2)$$

The phase is defined by

$$\tau(x,\phi) = x\cos\phi + \int_x^L \alpha(\xi)d\xi \, . \qquad (2.3)$$

The parameters β and δ are related, for TE polarization, by the equation

$$\tan(\beta\delta) = \{(n^2-1)\frac{1}{\beta^2} - 1\}^{1/2} \quad , \quad \delta \geqq 0. \qquad (2.4)$$

We now introduce the dependence of the parameters on x by letting the propagation constant in the x direction, vary linearly with x,

$$\alpha(x) = 1 + (\bar{\alpha} - 1)\frac{x}{L} \quad , \quad 0 \leqq x \leqq L. \qquad (2.5)$$

In (2.5) dielectric material properties determine the value of the constant

$$\bar{\alpha} \equiv \alpha(L) = 2.89 \, . \qquad (2.6)$$

By using (2.5) in (2.1) and (2,4) the parameters β and δ are also defined uniquely in terms of x. The dependence on x can be specified in ways other than by (2.5), for example by letting $\bar{\delta} = \delta x/L$. We will not consider these various possibilities here.

F. Schwering, using a local made theory [1], has obtained a closed form integral expression for the antenna radiation strength. In this expression the integrand is highly oscillatory, making quadrature difficult and expensive. His equation is

$$Q_A(\phi) = c\int_0^L q_1 q_2 \exp(-i\tau x)dx \, . \qquad (2.7)$$

Here the constant

$$c = \frac{n^2(n^2-1)}{2\sqrt{2\pi}} \{\frac{\bar{\alpha}(1 + \bar{\delta}\sqrt{\bar{\alpha}^2 - 1})}{(\bar{\alpha}^2 - 1)^{1/2}}\}^{1/2} \, , \qquad (2.8)$$

$$q_1 = \{\alpha^{-1}(\alpha^2 - 1)^{1/2}[1 + \delta(\alpha^2 - 1)^{1/2}]^{-1}\}^{1/2} \, , \qquad (2.9)$$

and

$$q_2 = \frac{\sin[(\beta + \sin\phi)\delta]}{\beta + \sin\phi} + \frac{\sin[(\beta - \sin\phi)\delta]}{\beta - \sin\phi} \, . \qquad (2.10)$$

In section three we will approximate (2.7) for long slender antennas,

$$(\text{ i.e. } L >> 1 ;\ \bar{\delta}/L << 1)$$

using asymptotic techniques to obtain a simplified analytic expression in the engineering variables. This expression is easily amenable to numerical calculation and at the same time exhibits the dependence of the antenna pattern on the physical variables. In (2.7) the functions q_1 and q_2 vary slowly throughout the region of intergration. However, the exponential varies rapidly because of the large parameter L.Hence (2.7) is an intergral with a highly oscillatory integrand. Therefore cancellation will occur in the integration process except in the neighborhood of special points - the stationary points and the end points of intergration. It is the contributions from the neighborhoods of these critical points which determine the value of the integral.

3. ASYMPTOTIC ANALYSIS - TE POLARIZATION

Our first step is to examine the range and variation of the parameters of the problem as x increases from zero to L. We see from (2.5) that α will increase monotonically from one to $\bar{\alpha}$. Consequently from (2.1) β will decrease monotonically from $\sqrt{11}$ to $\bar{\beta} = (12 - \bar{\alpha}^2)^{1/2}$. Then from (2.4) it follows that δ will increase monotonically from zero to a finite value , $\bar{\delta}$, since β occurs in the denominator of this expression. We will require in our analysis a more detailed behavior of δ as x approaches zero. From (2.1) and (2.4)

$$\delta = \frac{1}{\beta} \arctan\{\frac{\alpha^2 - 1}{\beta^2}\}^{1/2} .$$

Then use (2.5). As $x \to 0$ one sees that

$$\delta \sim (\alpha + 1)^{1/2}[(\bar{\alpha} - 1) \frac{x}{L}]^{1/2}/\beta^2 .$$

Finally the above becomes

$$\text{As } x \to 0; \quad \left[\begin{array}{l} \alpha \to 1 \\ \beta \to \sqrt{11} \\ \delta \cong \{2[\bar{\alpha} - 1]\}^{1/2}(\frac{x}{L})^{1/2}/11 \to 0. \end{array}\right. \tag{3.1}$$

Inserting (2.5) into (2.3) and integrating gives

$$\tau = - LH(s) , \tag{3.2}$$

where

$$H(s) = \frac{\bar{\alpha} - 1}{2} s^2 + (1 - \cos\phi)s - (\frac{\bar{\alpha} + 1}{2}) , \tag{3.3}$$

and

$$s = x/L , \quad 0 \leqq s \leqq 1. \tag{3.4}$$

Inserting (3.2) into (2.7) and making the change of variables of integration given by (3.4) we obtain

$$Q_A = LC \int_0^1 q_1 q_2 \exp\{iLH\} ds . \tag{3.5}$$

From (3.3)

$$H'(s) = (\bar{\alpha} - 1)s + (1 - \cos\phi) , \quad 0 \leqq s, \frac{\phi}{\pi} \leqq 1 , \qquad (3.6)$$

$$H''(s) = (\bar{\alpha} - 1) > 0 . \qquad (3,7)$$

From (3.6) we have two cases -

Case 1 : If $\phi = 0$ and $s = 0$ then $H'(0) = 0$. Thus there is a stationary point at the zero endpoint of integration for forward end-on radiation.

Case 2 : If $\phi \neq 0$ then $H'(s) \neq 0$ for all x. No stationary point occurs in the region of integration for all tnese radiation directions. For L much greater than one both cases can be handled asymtotically by the following theorem, as presented by Erdelyi [2].

THEOREM: For

$$\int_a^b g(s)\exp\{iLH(s)\}(s - a)^{\lambda-1}(b - s)^{\mu-1}ds = B_N(L) - A_n(L) , \qquad (3.8)$$

if

$$0 < \lambda,\mu \leqq 1; \quad g(s) \in C^N[a,b] ,$$

H(s) is differentiable, and

$$H'(s) = (s - a)^{\rho-1}(b - s)^{\sigma-1}H_1(s); \quad \rho,\sigma \geqq 1 \qquad (3.9)$$

$H_1(s) > 0$ and $\in C^N[a,b]$, then

$$A_N(L) = - \sum_{n=0}^{N-1} \frac{K^{(n)}(0)}{n!\rho} \Gamma(\frac{n + \lambda}{\rho})\exp[\frac{\pi i(n + \lambda)}{2\rho}]L^{-(\frac{n + 2}{2})} \exp[iLH(a)] \qquad (3.10)$$

$$U^\rho = H(s) - H(a) \qquad (3.11)$$

$$g_1(s) = g(s)(s - a)^{\lambda-1}(b - s)^{\mu-1} \qquad (3.12)$$

$$K(U) = g_1(s)U^{1-\lambda} \frac{ds}{dU} , \qquad (3.13)$$

and a similar expression for $B_N(L)$.

The above formulas handle not only simple end-point and stationary point contributions to the value of the integral but also include integrable algebraic singularities at the end points $(0 < \lambda,\mu < 1)$ and stationary points of higher order $(\rho,\sigma > 1)$.

In both cases we will omit the contribution from the upper endpoint, s = 1, as this will cancel with the contribution from the transmission line. We will now consider the two cases separately.

Case 1: Stationary Endpoint

Since

$$\phi = 0 \tag{3.14}$$

we have from (3.3) and (3.6)

$$H(s) = \frac{(\bar{\alpha} - 1)}{2} s^2 - \frac{(\bar{\alpha} + 1)}{2} ; \quad H(0) = -\frac{(\bar{\alpha} + 1)}{2} , \tag{3.15}$$

$$H'(s) = (\bar{\alpha} - 1)s , \quad H'(0) . \tag{3.16}$$

Applying (3.9) and (3.11) to (3.15) - (3.16) gives

$$\rho = 2 , \sigma = 1 \tag{3.17}$$

$$U^2 = \frac{(\bar{\alpha} - 1)}{2} s^2 . \tag{3.18}$$

Therefore

$$s = \left(\frac{2}{\bar{\alpha} - 1}\right)^{1/2} U \tag{3.19}$$

$$\frac{ds}{dU} = \left(\frac{2}{\bar{\alpha} - 1}\right)^{1/2} . \tag{3.20}$$

Since we are determining the contribution to the integral in the neighborhood of $s = 0$ we shall keep the firstterm of the product $q_1 q_2$ expanded in powers of s.

Applying (3.1) and (3.4) to (2.5), (2.9), and (2.10) we obtain as $s \to 0$

$$q_1 \cong \{2(\alpha - 1)\}^{1/4} = \{2(\bar{\alpha} - 1)\}^{1/4} s^{1/4} . \tag{3.21}$$

For $\phi \equiv 0$

$$q_2 = \frac{2}{\beta} \sin(\beta\delta) \cong 2\delta ,$$

$$q_2 = \frac{2^{3/2}}{11} (\bar{\alpha} - 1)^{1/2} s^{1/2} . \tag{3.22}$$

Therefore to the smallest power in s

$$q_1 q_2 \cong 2^{7/4} (\bar{\alpha} - 1)^{3/4} s^{3/4} / 11 . \tag{3.23}$$

Inserting (3.23) into (3.5) gives

$$Q_A(\phi=0) = 2^{7/4} (\bar{\alpha} - 1)^{3/4} (11)^{-1} CLR . \tag{3.24}$$

We write R in a special form for the application of the theorem as follows:

$$R = \int_{0+} s^{+1} \exp\{iLH\} s^{-1/4} ds . \tag{3.25}$$

Comparing (3.25) with (3.8) and (3.12) we see that

$$q_1 = s^{3/4} \tag{3.26}$$

$$\lambda = 3/4, \quad \mu = 1 . \tag{3.27}$$

Inserting (3.26), (3.27), and (3.20) in (3.13) gives

$$K(U) = s^{3/4}U^{1/4}(\frac{2}{\bar{\alpha} - 1})^{1/2} .$$

Substituting (3.19) into the above and simplifying yields

$$K(U) = (\frac{2}{\bar{\alpha} - 1})^{7/8}U, \quad K(0) = 0 . \tag{3.28}$$

Then

$$K'(0) = (\frac{2}{\bar{\alpha} - 1})^{7/8} \tag{3.29}$$

$$K^{(n)}(0) = 0 \ , \ n = 2,3,\cdots . \tag{3.30}$$

Inserting (3.28) - (3.30), (3.17), (3.27), and (3.15) into (3.10) yields the (n = 1) term, all other terms being zero,

$$R \cong - A_N = \frac{\Gamma(7/8)}{2^{1/8}(\bar{\alpha} - 1)^{7/8}} \exp\{i[\frac{7\pi}{6} - (\frac{\bar{\alpha} + 1}{2})L]\}L^{-7/8} . \tag{3.31}$$

Then from (3.24) and (3.31)

$$Q_A(\phi = 0) \cong \frac{2^{13/8}C\Gamma(7/8)}{(11)(\bar{\alpha} - 1)^{1/8}} \exp\{i[\frac{7\pi}{6} - (\frac{\bar{\alpha} + 1)}{2})L]\}L^{1/8} . \tag{3.32}$$

In the above Γ is the standard Gamma function and C is the constant given by (2.8). It is the formula we have been seeking. Note that

$$Q_A(\phi = 0) = O(L^{1/8}) . \tag{3.33}$$

Thus as the length of the antenna increases without limit, the the energy radiated in the forward direction of the antenna axis also increases without limit.

Case 2: Nonstationary Endpoint

Here

$$\phi \neq 0 . \tag{3.34}$$

Then H(s) and H'(s) are given by (3.3) and (3.6) respectively. Using these equations in (3.9) and (3.11) we obtain

$$\rho = 1, \quad \sigma = 1, \tag{3.35}$$

$$U' \equiv H(s) - H(0) = \frac{(\bar{\alpha} - 1)}{2}s^2 + (1 - \cos\phi)s \tag{3.36}$$

Then

$$\frac{dU}{ds} = (\bar{\alpha} - 1)s + (1 - \cos\phi) . \tag{3.37}$$

As s approaches one the approximation of q_1q_2 is unaffected by the value of ϕ. This is so since q_1 does not depend on ϕ and $q_2 \sim 2\delta$, as can be seen from (2.9), (2.10), and (3.1).

Therefore the equations (3.21) thru (3.27) remain equally valid for the case $\phi \neq 0$, where however H(s) is given by (3.3). Thus we can substitute (3.26), (3.36), (3.27) and (3.37) into (3.13) obtaining

$$K(U) = \frac{s[\frac{\bar{\alpha} - 1}{2} s + (1 - \cos\phi)]^{1/4}}{(\bar{\alpha} - 1)s + 1 - \cos\phi} \quad . \tag{3.38}$$

We need to calculate the first non-zero value of $K^{(n)}(0)$. Using (3.36) it follows that U and s are simultaneously zero. Therefore from (3.38)

$$K(0) = 0 \; . \tag{3.39}$$

Since

$$K^{(1)}(0) = \frac{dK}{ds} / \frac{dU}{ds} \bigg|_{s = 0} ,$$

we differentiate (3.38) with respect to s, divide by (3.37) and set s = 0. We obtain after simplification

$$K^{(1)}(0) = (1 - \cos\phi)^{-7/4}. \tag{3.40}$$

All the requisite quantities have now been computed.
We substitute (3.35), (3.27), (3.3), (3.39), and (3.40) into (3.10) obtaining to the highest order in L

$$R(\phi \neq 0) \cong -A_N = \frac{\Gamma(7/4)\exp\{i[\frac{7\pi}{8} - \frac{(\bar{\alpha} + 1)}{2} L]\}L^{-7/4}}{(1 - \cos\phi)^{7/4}} \quad . \tag{3.41}$$

Substituting (3.41) into (3.24) gives

$$Q(\phi \neq 0) \cong \frac{2^{7/4}(\bar{\alpha} - 1)^{3/4} C\Gamma(7/4)\exp\{i[\frac{7\pi}{8} - \frac{(\bar{\alpha} + 1)}{2} L]\}}{(11)(1 - \cos\phi)^{7/4}} L^{-3/4}. \tag{3.42}$$

This is the asymptotic expression we have been seeking. Note that

$$Q_A(\phi \neq 0) = \frac{1}{(1 - \cos\phi)^{7/4}} O(L^{-3/4}) \; . \tag{3.43}$$

Therefore as the length of the antenna increases, the radiation energy in all but the forward direction becomes vanishingly small. We also note that as ϕ approaches zero the asymptotic approximation (3.43) becomes infinite. Therefore, considered as a function of ϕ, it does not transform continuously into the asymptotic approximation (3.33). This is an asymptotic anomaly, or mathematical defect of the approximation method we have used, since the original integral (2.7), depends continuously on ϕ. A next step would be to determine a single asymptotic approximation valid uniformly with aspect to radiation direction.

4. FINAL COMMENTS

We wish to make two brief comments. On the basis of the foregoing analysis it is clear that as the length of the antenna increases the radiation pattern will become more highly directional. The field values can be calculated from (3.32) and (3.42) once the parameters $\bar{\alpha}$ and L have been specified. Secondly, an exactly analogous analysis can be performed when the antenna is operated in the TM (transverse magnetic) polarization mode. We do not perform this analysis here in order to keep the presentation relatively uncluttered and because no new mathematics is introduced.

REFERENCES

1. Personal communication from Felix Schwering.

2. A Erdelyi, Asymptotic Expansions, Dover Publications, 1956, pp 52-56.

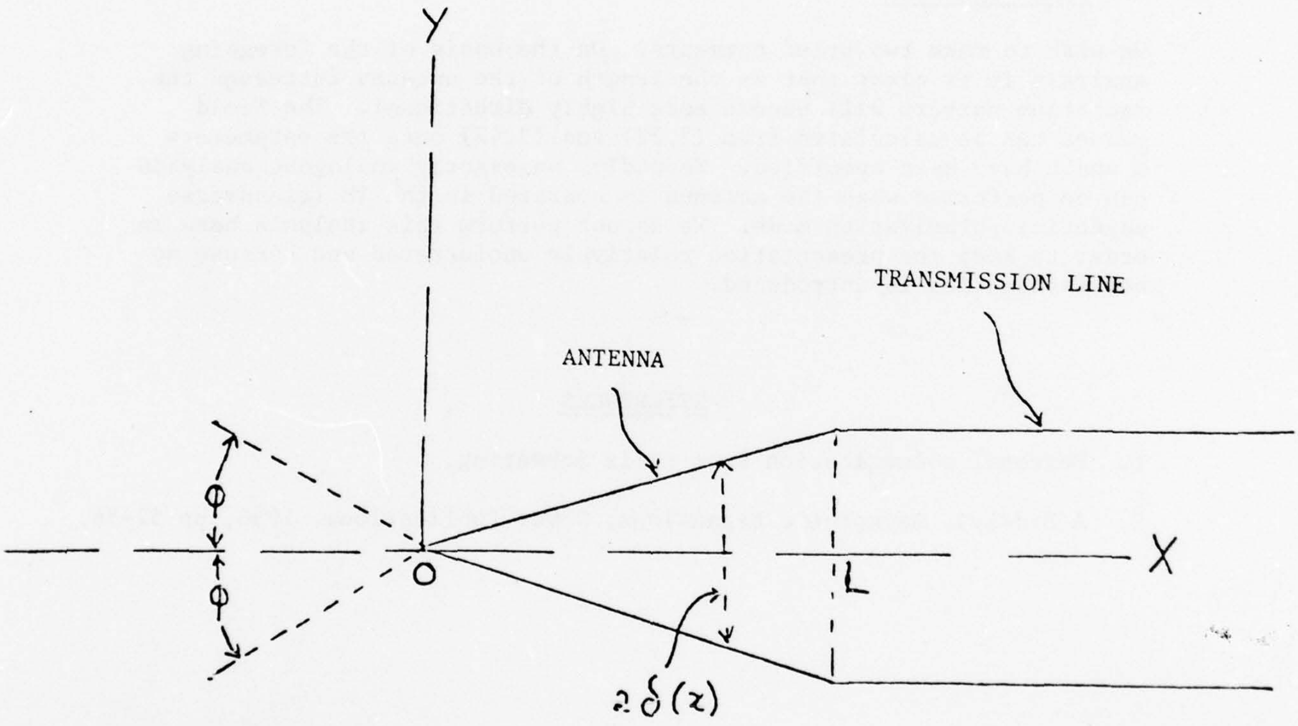


Figure 1. Two-dimensional Cross-section of Dielectric Antenna

THE INTEGRAL EQUATION OF IMAGE RECONSTRUCTION

L. B. Rall
Mathematics Research Center
University of Wisconsin-Madison
Madison, Wisconsin 53706

ABSTRACT. The important problem of reconstructing three-dimensional images from one- and two-dimensional data leads to an integral equation of Abel type. Some theoretical and practical aspects of the solution of this equation are discussed.

I. IMAGE RECONSTRUCTION. The general problem of determination of the internal structure of a system from external observations arises in many practical situations. For example, the strength of a material or a welded joint depends to some extent on the distribution of included particles or voids in size and number. Traditional methods of investigation include polishing the face of a slice of the material for microscopic examination, or the use of X-ray photographs to arrive at an estimate of the distribution parameters. This determination of three-dimensional structure from two- or one-dimensional sections or projections is the core topic of the subject of stereology [10] (which also deals with the reconstruction of two-dimensional images from one-dimensional data). There are a number of important subfields of stereology, such as seismology [4], which is concerned with the deduction of the internal structure of the earth from measurements made at the surface of pressure and shear waves generated by earthquakes or explosions, and tomography [9], which deals with the location of tumors in soft tissues of the human body by X-ray or microwave scanning.

Anderssen [2] lists a large number of stereological and other problems which lead to an integral equation of the form

$$\int_y^\infty k(y,x)(x^p - y^p)^{-\alpha} u(x)dx = s(y) , \qquad 0 < \alpha < 1 , \quad p > 0 , \tag{1}$$

where $s(y)$ is given for $y \geq 0$ and $u(x)$ is to be determined. The kernel $k(y,x)$ of the integral equation is assumed to be known and continuous; in most applications it is separable, that is

$$k(y,x) = k_1(y)k_2(x), \qquad 0 \leq y \leq x . \tag{2}$$

In the technical terminology of the theory of integral equations, (1) is called a Volterra integral equation of first kind with weakly singular kernel; more specifically, for $k(y,y) \neq 0$, it is called an Abel-type integral equation of first kind after the Norwegian mathematician who derived and solved a special case in connection with a mechanical problem more than 150 years ago [1]. Abel's problem will be discussed here in sections IV and V. Because of its importance

Sponsored by the United States Army under Contract No. DAAG29-75-C-0024.

in stereology, the integral equation (1) with kernel (2) will also be referred to as the integral equation of image reconstruction. Much use is made in what follows of the report by Anderssen [2], which also gives an extensive bibliography of works dealing with the theory and applications of this equation; only a few key and additional references will be cited here.

II. THE RANDOM SPHERES PROBLEM. Suppose one has a region in which spherical (or approximately spherical) particles are distributed at random. As examples, one can think of carbon particles in steel, or holes in Swiss cheese. A slice of such a material will show a random pattern of circles corresponding to the spheres cut by the sectioning plane (see Figure 1). The radii of the circles will vary due to the fact that the spheres are of different sizes, and are generally cut at some latitude other than their equator. If m is the average radius of the spheres and $s(y)$ is the density function of the distribution of radii of circles, then the unknown density function $u(x)$ of the radii of the spheres is given by the integral equation of image reconstruction with

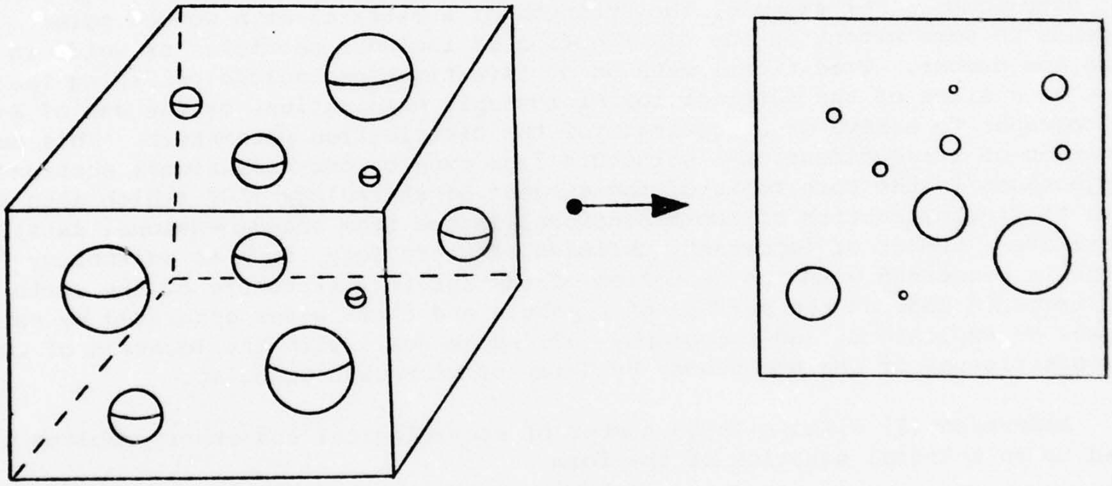


Figure 1. The Random Spheres Model.

$$k_1(y) = y/m, \quad k_2(x) = 1, \quad \alpha = \frac{1}{2}, \quad p = 2 \tag{3}$$

that is,

$$\int_y^\infty \frac{u(x)dx}{(x^2 - y^2)^{1/2}} = \frac{m}{y} s(y) \; . \tag{4}$$

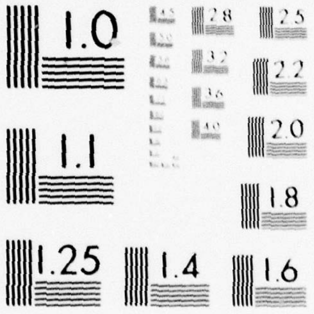
An elegant derivation of this equation is given in Chapter 16 of the book by Santaló [8]; see also [2, §3]. The quantity m can be found to be

$$m = \frac{\pi}{2} \left[\int_0^a \frac{s(x)}{x} dx\right]^{-1} \tag{5}$$

AD-A064 433   ARMY RESEARCH OFFICE RESEARCH TRIANGLE PARK N C   F/G 12/1
TRANSACTIONS OF THE CONFERENCE OF ARMY MATHEMATICIANS(24TH) HEL--ETC(U)
JAN 79

UNCLASSIFIED   ARO-79-1   NL

4 OF 6
AD
A064433



MICROCOPY RESOLUTION TEST CHART

NATIONAL BUREAU

By an independent argument [2, §5; 12], the upper limit a is the maximum radius of the circles. Another expression for m is

$$m = \int_0^\infty xu(x)dx \ , \tag{6}$$

assuming that u(x) is known.

Some theoretical and practical aspects of the solution of integral equations of Abel type and first kind, such as (4), will be given in the following sections.

III. SOLUTION BY INVERSION FORMULAS. The transformation of equation (1) with kernel given by (2) into an equation essentially of the form

$$\int_y^\infty k_1(y)k_2(x)(x-y)^{-\alpha}u(x)dx = s(y), \qquad 0 < \alpha < 1, \quad y \geq 0 \tag{7}$$

does not present any inherent difficulty, so the analysis will be carried out for this equation. The procedure will be formal; for justification, the appropriate conditions can be found in the paper by Atkinson [3]. First, write (7) as

$$\int_y^\infty \frac{k_2(t)u(t)dt}{(t-y)^\alpha} = \frac{s(y)}{k_1(y)} \ . \tag{8}$$

Now, the technique discovered long ago by Abel [1] may be applied: Multiply both sides of (8) by $(y - x)^{\alpha-1}$ and integrate with respect to y from x to ∞ to obtain

$$\int_x^\infty \frac{dy}{(y-x)^{1-\alpha}} \int_y^\infty \frac{k_2(t)u(t)dt}{(t-y)^\alpha} = \int_x^\infty \frac{s(y)dy}{k_1(y)(y-x)^{1-\alpha}} \ . \tag{9}$$

Interchange of the order of integration in (9) gives

$$\int_x^\infty k_2(t)u(t)dt \int_x^t \frac{dy}{(y-x)^{1-\alpha}(t-y)^\alpha} = \int_x^\infty \frac{s(y)dy}{k_1(y)(y-x)^{1-\alpha}} \ . \tag{10}$$

The second integral on the left can be evaluated explicitly; the change of variables $y = x + \theta(t - x)$ yields

$$\int_x^t \frac{dy}{(y-x)^{1-\alpha}(t-y)^\alpha} = \int_0^1 \frac{d\theta}{\theta^{1-\alpha}(1-\theta)^\alpha} \ , \tag{11}$$

where the integral on the right can be identified immediately as the Euler Beta-function [7, p. 95],

$$B(\alpha, 1-\alpha) = \Gamma(\alpha)\Gamma(1-\alpha) \ . \tag{12}$$

The further change of variables $\theta = (1 + w)^{-1}$ gives

$$B(\alpha, 1 - \alpha) = \int_0^\infty \frac{w^{-\alpha}dw}{1 + w}, \tag{13}$$

which can be evaluated by residues [5, pp. 131-133] to obtain finally

$$\int_x^t \frac{dy}{(y - x)^{1-\alpha}(t - y)^\alpha} = B(\alpha,\ 1 - \alpha) = \frac{\pi}{\sin\alpha\pi}. \tag{14}$$

Thus, from (10),

$$\int_x^\infty k_2(t)u(t)dt = \frac{\sin\alpha\pi}{\pi}\int_x^\infty \frac{s(y)dy}{k_1(y)(y - x)^{1-\alpha}}. \tag{15}$$

Differentiating (15) with respect to x gives the <u>first inversion formula</u>

$$u(x) = -\frac{\sin\alpha\pi}{\pi k_2(x)}\frac{d}{dx}\int_x^\infty \frac{s(y)dy}{(y - x)^{1-\alpha}k_1(y)}. \tag{16}$$

The *integrand in (16) is not smooth enough to* apply Leibniz' rule for differentiation under the integral sign; however, if $s(y)/k_1(y)$ is differentiable, one may integrate by parts to obtain

$$\int_x^\infty \frac{s(y)dy}{(y - x)^{1-\alpha}k_1(y)} = \lim_{y\to\infty}\left[\frac{(y - x)^\alpha}{\alpha u}\frac{s(y)}{k_1(y)}\right] - \int_x^\infty \frac{(y - x)^\alpha}{\alpha u}\frac{d}{dy}\left\{\frac{s(y)}{k_1(y)}\right\}dy. \tag{17}$$

Differentiation of (17) with respect to x and substitution into (16) results in the <u>second inversion formula</u>

$$u(x) = -\frac{\sin\alpha\pi}{\pi k_2(x)}\left[\lim_{y\to\infty}\left\{\frac{s(y)}{(y - x)^{1-\alpha}k_1(y)}\right\} + \int_x^\infty \frac{1}{(y - x)^{1-\alpha}}\frac{d}{dy}\left\{\frac{s(y)}{k_1(y)}\right\}dy\right]. \tag{18}$$

From a theoretical point of view, the inversion formulas (16) and (18) provide a complete solution of the integral equation (7), aside from the entertainment afforded by verification of the conditions under which various integrals exist and interchanges of limiting processes are valid. In practice, however, these formulas leave much to be desired. It is obvious from (18) that a differentiation of the data is required if one uses this formula for the solution of (7). *What if the data are* known only at discrete points, and are contaminated by noise? Then, one must perform a numerical differentiation of the data, which is a notoriously ill-posed problem. Because of the smoothing provided by the integration, the situation is not quite as bad if (16) is used; the extent of the necessary smoothness of the data will be clarified later. For this reason, the solution of Abel equations of the first kind is called a "weakly ill-posed" problem by Anderssen [2]. The basic features of the nature of these difficulties will be shown in the next two sections using as a paradigm the old mechanical problem solved by Abel.

IV. ABEL'S MECHANICAL PROBLEM. The determination of the path along which a body will slide a vertical distance h in specified time $T(h)$ is a problem which apparently goes back to Christiaan Huygens, a contemporary of Newton [6, pp. 246-247]. The body starts from rest, and the sliding is supposed to be frictionless (see Figure 2). At $T = 0$, the body has potential energy mgh, referred to h, where m is its mass. By the time t the particle is at a vertical distance y above h, the amount of potential energy converted into kinetic energy is

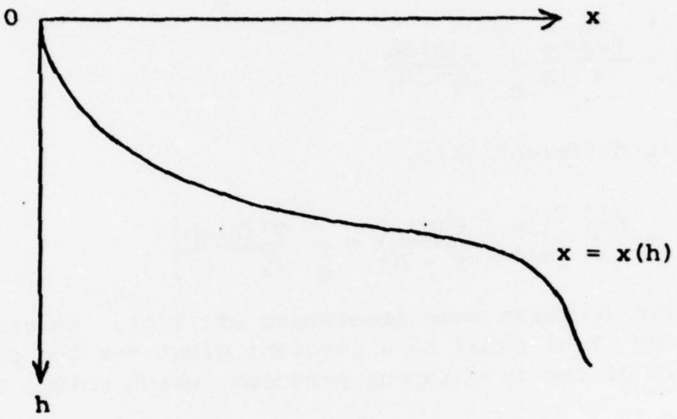


Figure 2. Abel's Mechanical Problem

$$\frac{1}{2} m\left(\frac{ds}{dt}\right)^2 = mg(h - y) \ , \tag{19}$$

where s denotes arc length along the path. Solving for dt and integrating,

$$\int_0^{S(h)} \frac{ds}{\sqrt{2g(h-y)}} = T(h) \ , \tag{20}$$

where $S(h)$ is the total length of the path traversed in descending a vertical distance h. Relating arc length s and y by

$$ds = -u(y)dy \ , \tag{21}$$

the result is the classical Abel equation

$$\int_0^h \frac{u(y)\,dy}{\sqrt{h - y}} = \sqrt{2g}\ T(h) \tag{22}$$

for $u(y)$, given $T(h)$. Once this function has been determined, the unknown path in Figure 2 is given by

$$x = \int_0^h \sqrt{[u(y)]^2 - 1}\, dy \ . \qquad (23)$$

Corresponding to the inversion formulas (16) and (18), the solution of (22) may be written as

$$u(y) = \frac{\sqrt{2g}}{\pi} \frac{d}{dy} \int_0^y \frac{T(h)\,dh}{\sqrt{y - h}} \ , \qquad (24)$$

or, if $T(h)$ is differentiable,

$$u(y) = \frac{\sqrt{2g}}{\pi} \left[\lim_{h \to 0} \left\{ \frac{T(h)}{\sqrt{y - h}} \right\} + \int_0^y \frac{T'(h)\,dh}{\sqrt{y - h}} \right] \ . \qquad (25)$$

Formula (24) also requires some smoothness of $T(h)$, as explained in the next section. Setting $T(h)$ equal to a constant gives for the path the famous inverted cycloid of the isochronous pendulum, which solves Huygen's problem of 1673 [6].

An image reconstruction problem connected with equation (22) is the following: Choose stations (points) along the h-axis, $0 < h_1 < h_2 < \ldots < h_n$, and measure the times $T_1, T_2, \ldots, T_n$ at which the body passes the corresponding station; from this data, find at least a good approximation to the path $x = x(h)$, taking into account that the given times are subject to observational errors. There are two approaches to this problem, one of which involves working directly with the original equation (22) and using some technique such as regularization, while the other would consist of constructing a sufficiently smooth approximation $T(h)$ to the data to permit the application of an inversion formula [2, §4]. Some aspects of the latter technique will now be discussed.

V. FRACTIONAL-ORDER DIFFERENTIATION. Motivated by the study of Abel integral equations, the derivative of order β of a function $f(x)$ can be defined by the formula

$$\frac{d^\beta}{dx^\beta} f(x) = \frac{1}{\Gamma(1-\beta)} \frac{d}{dx} \int_0^x \frac{f(t)\,dt}{(x-t)^\beta} \ , \qquad 0 < \beta < 1 \ . \qquad (26)$$

It turns out that this definition is also consistent at $\beta = 0$, giving $d^0 f(x)/dx^0 = f(x)$, and at $\beta = 1$, where a limiting process gives a version of the Cauchy integral formula for $f(x)$ if the limit of the ratio of the integral to the Gamma-function is taken first, followed by the differentiation to obtain $f'(x)$. Actually, (26) can be extended to all real numbers β, with negative values indicating integration, provided suitable assumptions about $f(x)$ are made at $x = 0$. In terms of this definition, the inversion formula for the Abel mechanical equation (22) becomes

$$u(y) = \sqrt{\frac{2g}{\pi}} \frac{d^{\frac{1}{2}}}{dy^{\frac{1}{2}}} T(y) , \tag{27}$$

Recalling that $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$. Thus, the use of the inversion formula requires a fractional differentiation of order $\frac{1}{2}$ of the data. Continuing this line of reasoning, the first inversion formula (16) for the general Abel equation (7) becomes

$$u(x) = - \frac{1}{k_2(x)\Gamma(1-\alpha)} \frac{d^{1-\alpha}}{dx^{1-\alpha}} \left\{\frac{s(y)}{k_1(y)}\right\} , \tag{28}$$

the fractional differentiation of the data being of order $1 - \alpha$.

Thus, in the reconstruction problem of the previous section, a piecewise linear interpolation of the data is not smooth enough. On the other hand, cubic spline interpolation, which results in continuous second derivatives, may be too smooth in that some interesting irregularities in the actual path will be lost. The weakly ill-posed nature of this problem induced by its fractional order as an integral equation also appears if equation (7) is attacked directly, as there is a close connection between the optimal regularization parameter and optimal smoothing (filtering) of the data [11].

VI. LINEAR FUNCTIONALS OF THE SOLUTION. In many practical problems, the difficulties brought on by the weak ill-posedness of equation (7) can be avoided if what one really wants is not its solution $u(x)$, but rather some linear functional

$$W_a[u] = \int_0^a w(x)u(x)dx , \tag{29}$$

with smooth kernel $w(x)$. For example, in the random spheres problem, one may want to know the fraction of spheres with radius less than or equal to a; here $w(x) \equiv 1$. For probability density functions $u(x)$, frequent choices of $w(x)$ would be $1, x, x^2, \ldots$ to obtain the moments of the corresponding distribution, and so on. If $w(x)$ is differentiable, then integration by parts may be applied to (29), using the first inversion formula (16), to obtain

$$W_a[u] = \left(- \frac{w(x)\sin\alpha\pi}{\pi k_2(x)}\right) \int_x^\infty \frac{s(y)dy}{(y-x)^{1-\alpha}k_1(y)} \Bigg|_{x=0}^{x=a} + \tag{30}$$

$$+ \int_0^a \int_x^\infty \frac{s(y)dy}{(y-x)^{1-\alpha}k_1(y)} \frac{d}{dx} \left\{\frac{w(x)\sin\alpha\pi}{\pi k_2(x)}\right\} dx .$$

Here, the integrals provide a little smoothing of the data, even if only of fractional order. Thus, it appears more sensible to use (30) in this circumstance in preference to solving (7) by some approximate method, and then substituting the result into (29) [2, §5].

REFERENCES.

1. N. H. Abel, Oeuvres Complètes, 2 vols., Christiania, Oslo, 1881; Vol. 1; pp. 11-27 (1823), pp. 97-101 (1826).

2. R. S. Anderssen, Application and numerical solution of Abel-type integral equations, MRC Technical Summary Report #1787, University of Wisconsin-Madison, 1977.

3. K. E. Atkinson, An existence theory for Abel integral equations, SIAM J. Math. Anal. 5 (1974), 729-736.

4. K. E. Bullen, An Introduction to the Theory of Seismology, Cambridge University Press, 1963.

5. R. V. Churchill, Introduction to Complex Variables and Applications, McGraw-Hill, New York, 1948.

6. A. T. Lonseth, Sources and applications of integral equations, SIAM Rev. 19 (1977), 241-278.

7. B. O. Peirce, A Short Table of Integrals, 4th Ed., Revised by Ronald M. Foster, Ginn and Company, Boston, 1956.

8. L. A. Santaló, Integral Geometry and Geometric Probability, Addison-Wesley, Reading, Massachusetts, 1976.

9. K. T. Smith, D. C. Solmon, and S. L. Wagner, Practical and mathematical aspects of the problem of reconstructing objects from radiographs, Bull. Amer. Math. Soc. 83 (1977), 1227-1270. Addendum, 84 (1978), 691.

10. E. E. Underwood, Quantitative Stereology, Addison-Wesley, Reading, Massachusetts, 1970.

11. G. Wahba, Practical approximate solutions to linear operator equations where the data are noisy, SIAM J. Numer. Anal. 14 (1977), 651-667.

12. G. S. Watson, Estimating Functionals of particle size distributions, Biometrika 58 (1971), 483-490.

STOCHASTIC INTEGRAL EQUATIONS

Marc A. Berger
Mathematics Research Center
University of Wisconsin at Madison
Madison, WI 53706

ABSTRACT. This paper summarizes some of the results in Berger [2], [3], [4] and Berger and Mizel [5] on stochastic integral equations. The presentation and the manipulations are intended to be formal, motivated entirely by examples. Compound noise effects and feedback in the presence of noise are illustrated. A theorem on changing the order of integration in a double stochastic integral, and a resolvent formula for linear stochastic Volterra equations are included.

I. INTRODUCTION. Let $(\Omega, F, \mathbb{P})$ be a probability space. White noise is generally taken to be a stationary Gaussian process $\{z(t) : t \in (-\infty,\infty)\}$ with mean identically zero, and a constant spectral density. It is easily seen that such a process does not exist in the classical sense, and the interested reader is referred to Gelfand and Vilenkin [6] for a rigorous description.

II. STOCHASTIC DIFFERENTIAL EQUATIONS. In studying the effect of white noise on the solutions of ordinary differential equations, one is led, through the stochastic integral, to the well-known subject of stochastic differential equations. (See, for example, Arnold [1].) Thus a Brownian motion $\{\beta(t) : t \geq 0\}$ is defined by $\beta(t) = \int_0^t z(\tau)d\tau$, and the ordinary differential equation

$$\text{(ODE)} \qquad \dot{x}(t) = a(t,x(t)) + \sigma(t,x(t))z(t)$$

is converted to the stochastic differential equation

$$\text{(SDE)} \qquad dx(t) = a(t,x(t))dt + \sigma(t,x(t))d\beta(t) .$$

For example, the Langevin equation

$$\dot{v}(t) = -av(t) + \sigma z(t) \qquad (a > 0, \sigma \text{ constants}) ,$$

describing the velocity of a tiny particle in fluid, corresponds to the linear stochastic differential equation

$$dv(t) = -av(t)dt + \sigma d\beta(t) .$$

An important consideration in the subject of stochastic differential equations is that of asymptotic stability. The one-dimensional linear homogeneous equation can be used to illustrate the types of results one desires. The solutions to the equation

$$dx(t) = ax(t)dt + \sigma x(t)d\beta(t) \qquad (a, \sigma \text{ constants})$$

Sponsored by the United States Army under Contract No. DAAG-29-75-C-0024 and the National Science Foundation under Grant No. MCS75-17385 A01.

are

$$x(t) = ce^{(a - \frac{1}{2}\sigma^2)t + \sigma\beta(t)} .$$

Thus the necessary and sufficient condition for a.s. asymptotic stability is $a < \frac{1}{2}\sigma^2$, and the necessary and sufficient condition for mean-square stability is $a \leq -\frac{1}{2}\sigma^2$. In particular, if the unperturbed equation $(\sigma = 0)$ is asymptotically stable then the perturbed equation is asymptotically stable a.s. For results and examples on multi-dimensional equations the reader is referred to Khasminski [8] and Pinsky [9].

III. STOCHASTIC INTEGRAL EQUATIONS. Just as in the deterministic case, many stochastic models can be more accurately described through hereditary-type equations. Instead of focusing on equations of the form (SDE), one can consider a more general stochastic differential equation

$$\text{(SDE)}' \qquad dx(t) = a(t,\{x(\tau) : 0 \leq \tau \leq t\})dt + \sigma(t,\{x(\tau) : 0 \leq \tau \leq t\})d\beta(t) .$$

For the measurability conditions, assume that the processes $a(t,\{f(\tau) : 0 \leq \tau \leq t\})$ and $\sigma(t,\{f(\tau) : 0 \leq \tau \leq t\})$ are non-anticipating when f is a deterministic function. For example,

$$\sigma(t,\{f(\tau) : 0 \leq \tau \leq t\}) = \int_0^t k(t,\tau)f(\tau)d\beta(\tau) ,$$

or

$$\sigma(t,\{f(\tau) : 0 \leq \tau \leq t\}) = \phi(t,f(t)) .$$

Of course, it is understood that the solutions of (SDE)' will not, in general, be Markov. However, questions concerning stationarity, blowing up of solutions, asymptotic behavior, etc. are still of interest. The case where the coefficients a and σ are independent of t appears in Ito and Nisio [7].

A related equation is

$$(*) \qquad x(t) = a(t,\{x(\tau) : 0 \leq \tau \leq t\}) ,$$

where a satisfies the same measurability condition as before. Under suitable differentiability assumptions on a, $(*)$ can be converted to an equation of the form (SDE)'. Thus

$$x(t) - \int_0^t k(t,\tau)x(\tau)d\beta(\tau) = f(t)$$

becomes

$$dx(t) = [\dot{f}(t) + \int_0^t k_t(t,\tau)x(\tau)d\beta(\tau)]dt + k(t,t)x(t)d\beta(t) .$$

IV. EXAMPLES. The first example concerns the circuit below. Thermal noise, acting on the current I(t), through the resistor R results in a dissipation

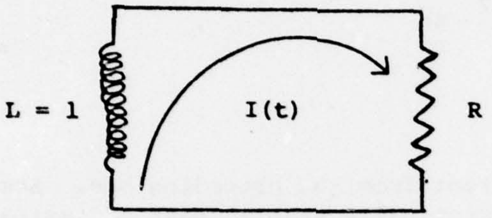
$$H(t) = \sigma I(t) z(t) \ . \qquad (4.1)$$

Assuming an inductance of one, the circuit equation is

$$\dot{I}(t) + RI(t) + H(t) = 0 \ ,$$

$$I(0) = I_0$$

This leads to the equation



$$dI(t) = -RI(t)dt - \sigma I(t) d\beta(t)$$

the solution of which is

$$I(t) = I_0 e^{-(R+\frac{1}{2}\sigma^2)t - \sigma\beta(t)} \ .$$

If the dissipation has a hereditary-type effect,

$$H(t) = [\sigma I(t) + \eta \int_0^t e^{-R(t-\tau)} I(\tau) z(\tau) d\tau] z(t) \ , \qquad (4.2)$$

then the circuit equation leads to the stochastic integro-differential equation

$$dI(t) = -RI(t)dt - [\sigma I(t) + \eta \int_0^t e^{-R(t-\tau)} I(\tau) d\beta(\tau)] d\beta(t) \ .$$

The solution of this equation is

$$I(t) = \begin{cases} \frac{1}{2} I_0 \left[\left(1 + \frac{\sigma}{\sqrt{\sigma^2 - 4\eta}}\right) e^{\lambda_1 t + \mu_1 \beta(t)} + \left(1 - \frac{\sigma}{\sqrt{\sigma^2 - 4\eta}}\right) e^{\lambda_2 t + \mu_2 \beta(t)} \right] \ , & \sigma^2 \neq 4\eta \ , \\ I_0 [1 - \eta t - \frac{1}{2}\sigma\beta(t)] e^{\lambda t + \mu\beta(t)} \ , & \sigma^2 = 4\eta \ , \end{cases}$$

where

$$\lambda_1 = -(R + \frac{1}{4}\sigma^2 - \frac{1}{2}\eta) - \frac{1}{4}\sigma\sqrt{\sigma^2 - 4\eta} ,$$

$$\lambda_2 = -(R + \frac{1}{4}\sigma^2 - \frac{1}{2}\eta) + \frac{1}{4}\sigma\sqrt{\sigma^2 - 4\eta} ,$$

$$\lambda = - R - \frac{1}{2}\eta ,$$

$$\mu_1 = -\frac{1}{2}\sigma - \frac{1}{2}\sqrt{\sigma^2 - 4\eta} ,$$

$$\mu_2 = -\frac{1}{2}\sigma + \frac{1}{2}\sqrt{\sigma^2 - 4\eta} ,$$

$$\mu = -\frac{1}{2}\sigma .$$

This case is entirely different from the preceding one. Assuming $R > 0$, the current is always a.s. asymptotically stable under a dissipation of the form (4.1), for any σ. But under a dissipation of the form (4.2) the current is a.s. asymptotically stable if and only if

$$\eta < 2R + \frac{1}{2}\sigma^2 .$$

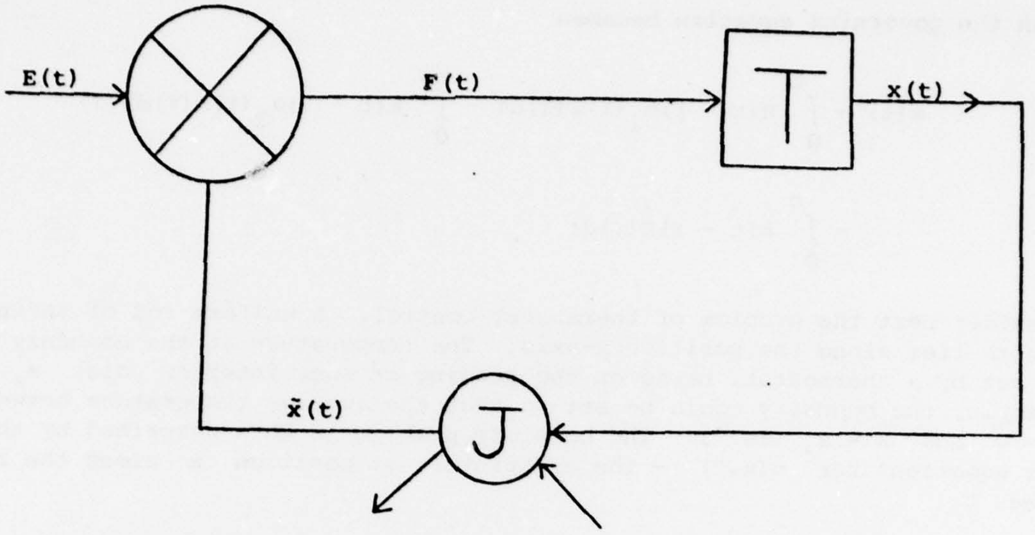
Thus, whereas σ can only have a stabilizing effect, the hereditary coefficient η can have both a stabilizing and a de-stabilizing effect. In Berger [4] it is shown that whenever the Fourier transform of $k(t)$ has a non-negative real part, the solutions of

$$dx(t) = ax(t)dt + \left[\int_0^t k(t - \tau)x(\tau)d\beta(\tau)\right]d\beta(t)$$

satisfy

$$|x(t)| \geq |x_0|e^{[a - \frac{1}{2}k(0)]t} , \quad \text{a.s.}$$

The next examples concern feedback in the presence of white noise. Shown in the figure is a typical feedback diagram. The box T signifies a transfer from the input F



to the output x. For example,

$$x(t) = \int_0^t k(t - \tau)F(\tau)d\tau \ . \tag{4.3}$$

Junction J is a step-up or step-down point. Here either some fraction of x is diverted for external consumption, or else $\dot{x}$ is scaled up. Thus the remainder in the loop is

$$\tilde{x} = \alpha x \ . \tag{4.4}$$

If the process uses this remainder $\tilde{x}$ to drive itself, along with an external driving force E, then

$$F = E + \tilde{x} \ . \tag{4.5}$$

Combining (4.3), (4.4), (4.5) it follows that the equation governing the system is

$$x(t) - \int_0^t k(t - \tau)\alpha(\tau)x(\tau)d\tau = \int_0^t k(t - \tau)E(\tau)d\tau \ .$$

Suppose, however, that α is in the form of a noise

$$\alpha = \alpha_1 + \alpha_2 z \ .$$

Then the governing equation becomes

$$x(t) - \int_0^t k(t-\tau)\alpha_1(\tau)x(\tau)d\tau - \int_0^t k(t-\tau)\alpha_2(\tau)x(\tau)d\beta(\tau)$$
$$= \int_0^t k(t-\tau)E(\tau)d\tau \; . \tag{4.6}$$

Consider next the problem of thermostat control. A uniform rod of infinite length lies along the positive x-axis. The temperature at the boundary $(x = 0)$ is set by a thermostat, based on the reading at some interior point x_*. For example, the boundary could be set so that the average temperature between $x = 0$ and $x = x_*$ is T. The complete problem is then described by the following equations for $u(x,t)$ - the temperature at position x along the rod at time t.

$$u_t = \gamma u_{xx} \tag{4.7}$$

$$\frac{1}{2}\,[u(0,t) + u(x_*,t)] = T \tag{4.8}$$

$$u(x,0) = u_0(x) \tag{4.9}$$

The procedure for solving these equations is straightforward. Let

$$\Gamma(x,y,t) = \frac{1}{2\sqrt{\gamma\pi t}}\left[e^{-\frac{(x-y)^2}{4\gamma t}} - e^{-\frac{(x+y)^2}{4\gamma t}}\right]$$

Then $v(t) \equiv u(x_*,t) - T$ is the solution of the Volterra equation

$$v(t) + \gamma\int_0^t \Gamma_y(x_*,0,t-\tau)v(\tau)d\tau = \int_0^\infty \Gamma(x_*,y,t)[u_0(y) - T]dy \; .$$

If the reading at $x = x_*$ is disturbed by white noise then (4.8) should properly be replaced by

$$\frac{1}{2}\,\{u(0,t) + [1 + \alpha(t)z(t)]u(x_*,t)\} = T \; ,$$

and then the Volterra equation for $v(t)$ becomes

$$v(t) + \gamma \int_0^t \Gamma_y(x_*,0,t - \tau)v(\tau)d\tau + \gamma \int_0^t \Gamma_y(x_*,0,t - \tau)\alpha(\tau)v(\tau)d\beta(\tau)$$

$$= \int_0^\infty \Gamma(x_*,y,t)[u_0(y) - T]dy - \gamma T \int_0^t \Gamma_y(x_*,0,t - \tau)\alpha(\tau)d\beta(\tau) \tag{4.10}$$

Equations (4.6) and (4.10) are examples of linear stochastic Volterra equations. The general linear equation is

$$x(t) - \int_0^t a(t,\tau)x(\tau)d\tau - \int_0^t b(t,\tau)x(\tau)d\beta(\tau) = g(t) ,$$

and whenever a has a resolvent this equation can be reduced to

$$x(t) - \int_0^t k(t,\tau)x(\tau)d\beta(\tau) = f(t) . \tag{4.11}$$

Indeed, let $a_*(t,\tau)$ be the solution of

$$a_*(t,\tau) - \int_\tau^t a(t,s)a_*(s,\tau)ds = a(t,\tau) .$$

Then

$$f(t) = g(t) + \int_0^t a_*(t,\tau)g(\tau)d\tau ,$$

$$k(t,\tau) = b(t,\tau) + \int_\tau^t a_*(t,s)b(s,\tau)ds .$$

Thus, in order to solve (4.6) and (4.10) it is enough to derive a resolvent formula for the solution of (4.11). Based on the classical analysis of integral equations (cf. Riesz-Nagy [10]) one would construct a resolvent kernel as follows. Define

$$k_1(t,\tau) = k(t,\tau) ,$$

$$k_{n+1}(t,\tau) = \int_\tau^t k(t,s)k_n(s,\tau)d\beta(s) ; \qquad n = 1,2,...,$$

$$k_*(t,\tau) = \sum_{n=1}^\infty k_n(t,\tau) .$$

That is, k_* is to be a solution of

$$k_*(t,\tau) - \int_\tau^t k(t,s)k_*(s,\tau)d\beta(s) = k(t,\tau) \ . \tag{4.12}$$

Then the solution x ought to be given by

$$x(t) = f(t) + \int_0^t k_*(t,\tau)f(\tau)d\beta(\tau) \ . \tag{4.13}$$

But there are two drawbacks with this argument. First of all, the integrand of the stochastic integral in (4.13) anticipates, since $k_*(t,\tau)$ depends on the Brownian path up to time t. Thus the stochastic integral in (4.13) needs to be defined. This can be done in an intuitive way. However, rather than go through the details of the general definition, which appear in Berger [3], an example can be used to illustrate the ideas. Let $k(t,\tau) \equiv 1$. Then, by Ito's Formula,

$$k_n(t,\tau) = \frac{1}{n-1} H_{n-1}(\beta(t) - \beta(\tau), t-\tau); \qquad n = 1,2,\ldots,$$

where $H_n(x,t)$ is the Hermite polynomial of degree n, defined as the solution of

$$(H_n)_t + \frac{1}{2}(H_n)_{xx} = 0 \ ,$$

$$H_n(x,0) = x^n$$

Thus

$$k_*(t,\tau) = e^{\beta(t)-\beta(\tau)-\frac{1}{2}(t-\tau)} \ ,$$

and, indeed, this is the solution of (4.12). So the definition of $\int_0^t k_*(t,\tau)f(\tau)d\beta(\tau)$ is simply

$$e^{\beta(t)-\frac{1}{2}t} \int_0^t e^{-\beta(\tau)+\frac{1}{2}\tau} f(\tau)d\beta(\tau) \ ,$$

and the first drawback in deriving (4.13) can be circumvented.

The second drawback is that (4.13) yields an incorrect solution of (4.11)! Again, let $k(t,\tau) \equiv 1$. Then if f is differentiable, (4.11) can be written as a stochastic differential equation.

$$dx(t) = x(t)d\beta(t) + \dot{f}(t)dt \ ,$$

$$x(0) = f(0) \ .$$

The solution x is therefore given by

$$x(t) = f(0)k_*(t,0) + \int_0^t k_*(t,\tau)\dot{f}(\tau)d\tau$$

$$= f(t) + \int_0^t k_*(t,\tau)f(\tau)d\beta(\tau) - \int_0^t k_*(t,\tau)f(\tau)d\tau \ ,$$

and this does not agree with (4.13).

To understand what went wrong it is enough to examine the first few terms of the Neumann expansion. The justification for (4.13) is based on successive approximations. Define

$$x_1(t) = f(t)$$

$$x_{n+1}(t) = f(t) + \int_0^t k(t,\tau)x_n(\tau)d\beta(\tau); \qquad n = 1,2,\ldots \ .$$

It is shown in Berger [3] that under mild conditions on k, this sequence converges to the unique solution x of (4.11). In the classical case it would also be true that

$$x_n(t) = f(t) + \sum_{j=1}^{n-1} \int_0^t k_j(t,\tau)f(\tau)d\beta(\tau) \ ; \qquad n = 1,2,\ldots \ .$$

This is based on Fubini's Theorem on changing the order of integration in a double integral. Thus, for example,

$$x_2(t) = f(t) + \int_0^t k(t,\tau)f(\tau)d\beta(\tau) + \int_0^t k(t,\tau)\left[\int_0^\tau k(\tau,s)f(s)d\beta(s)\right]d\beta(\tau)$$

$$= f(t) + \int_0^t k(t,\tau)f(\tau)d\beta(\tau) + \int_0^t \left[\int_\tau^t k(t,s)k(s,\tau)d\beta(s)\right]f(\tau)d\beta(\tau) \ .$$

However, here lies the discrepency between the stochastic and the classical case. Fubini's Theorem does not hold in the stochastic case! Thus it is <u>not</u> true that

$$\int_0^t \int_0^\tau \phi(\tau,s)d\beta(s)d\beta(\tau) = \int_0^t \int_s^t \phi(\tau,s)d\beta(\tau)d\beta(s) \ .$$

This is easily seen, for if $\phi(\tau,s) \equiv 1$ then

$$\int_0^t \int_0^\tau \phi(\tau,s)d\beta(s)d\beta(\tau) = \int_0^t \beta(\tau)d\beta(\tau) = \frac{1}{2}\beta^2(t) - \frac{1}{2}t \ ,$$

and

$$\int_0^t \int_s^t \phi(\tau,s)d\beta(\tau)d\beta(s) = \int_0^t \big[\beta(t) - \beta(s)\big]d\beta(s)$$

$$= \beta^2(t) - \int_0^t \beta(s)d\beta(s) = \frac{1}{2}\beta^2(t) + \frac{1}{2}t \ .$$

In Berger [2], [3] it is shown that for a wide class of two-parameter stochastic processes $\{\phi(\tau,s) : 0 \leq s \leq \tau\}$ the following result is true:

$$\int_0^t \int_s^t \phi(\tau,s)d\beta(\tau)d\beta(s) = \int_0^t \int_0^\tau \phi(\tau,s)d\beta(s)d\beta(\tau) + \int_0^t \phi(\tau,\tau)d\tau \ . \quad (4.14)$$

For example, if $\phi(\tau,s) \equiv 1$ then the difference between the two double integrals is t, which is consistent with the above calculations. As another example, if $\phi(\tau,s) = \beta(\tau)\beta(s)$ then it follows from Ito's Formula that

$$\int_0^t \int_0^\tau \phi(\tau,s)d\beta(s)d\beta(\tau) = \frac{1}{8}[\beta^2(t) - t]^2 - \frac{1}{2}\int_0^t \beta^2(\tau)d\tau \ ,$$

$$\int_0^t \int_s^t \phi(\tau,s)d\beta(\tau)d\beta(s) = \frac{1}{8}[\beta^2(t) - t]^2 + \frac{1}{2}\int_0^t \beta^2(\tau)d\tau \ .$$

Now using (4.14) it can be seen by induction that in fact for $n \geq 2$

$$x_n(t) = f(t) + \sum_{j=1}^{n-1} \int_0^t k_j(t,\tau)f(\tau)d\beta(\tau) - \sum_{j=1}^{n-2} \int_0^t k_j(t,\tau)k(\tau,\tau)f(\tau)d\tau \ . \quad (4.15)$$

Indeed, for $n = 2$ the result is clear. Furthermore, if it is true for $n = m$ then

$$x_{m+1}(t) = f(t) + \int_0^t k(t,\tau)x_m(\tau)d\beta(\tau)$$

$$= f(t) + \sum_{j=1}^{m} \int_0^t k_j(t,\tau)f(\tau)d\beta(\tau) - \sum_{j=2}^{m-1} \int_0^t k_j(t,\tau)k(\tau,\tau)f(\tau)d\tau$$

$$-\int_0^t k(t,\tau)k(\tau,\tau)f(\tau)d\tau \ ,$$

since $k_j(\tau,\tau) \equiv 0$ for $j \geq 2$. Thus (4.15) is established, and the solution of (4.11) is given by

$$x(t) = f(t) + \int_0^t k_*(t,\tau)f(\tau)d\beta(\tau) - \int_0^t k_*(t,\tau)k(\tau,\tau)f(\tau)d\tau \ . \qquad (4.16)$$

This is consistent with the example $k(t,\tau) \equiv 1$ above. As another example, let $k(t,\tau) = k_1(t)k_2(\tau)$. Then, again using Ito's Formula, it follows that

$$k_*(t,\tau) = k(t,\tau)e^{\int_\tau^t k(s,s)d\beta(s) - \frac{1}{2}\int_\tau^t k^2(s,s)ds} \ ,$$

and x is given by (4.16). If k_1 and f are differentiable, this result can be checked directly by converting (4.11) into a stochastic differential equation. Namely,

$$dx(t) = \left\{\frac{\dot{k}_1(t)}{k_1(t)}[x(t) - f(t)] + \dot{f}(t)\right\}dt + k(t,t)x(t)d\beta(t) \ ,$$

$$x(0) = f(0)$$

For proofs and additional results, examples and applications the reader is referred to Berger [2], [3], [4].

REFERENCES.

1. Arnold, L., Stochastic Differential Equations: Theory and Applications, New York, Wiley-Interscience, 1974.

2. Berger, M.A., "Stochastic Ito-Volterra Equations", Ph.D. Dissertation, Carnegie-Mellon University, Pittsburgh, February, 1977.

3. Berger, M.A., "A Fubini Theorem for Iterated Stochastic Integrals", MRC Technical Summary Report #1826, February, 1978.

4. Berger, M.A., "Positive Kernels and Stochastic Integrals", MRC Technical Summary Report #1831, February, 1978.

5. Berger, M.A. and Mizel, V.J., "A Fubini Theorem for Iterated Stochastic Integrals", Bulletin of the American Mathematical Society, Vol. 84, pp. 159-160 (1978).

6. Gelfand, I.M. and Vilenkin, N.J., Generalized Functions, Vol. 4, New York, Academic Press, 1961 (translation from Russian).

7. Ito, K. and Nisio, M., "On Stationary Solutions of a Stochastic Differential Equation", Journal of Mathematics of Kyoto University, Vol. 4, pp. 1-75 (1964).

8. Khasminski, R.Z., "Necessary and Sufficient Conditions for the Asymptotic Stability of Linear Stochastic Systems", Theory of Probability and Applications, Vol. 12, pp. 144-147 (1967).

9. Pinsky, M.A., "Stochastic Stability and the Dirichlet Problem", Communications on Pure and Applied Mathematics, Vol. 27, pp. 311-350 (1974).

10. Riesz, F. and Nagy, B.S., Functional Analysis, New York, Frederick Ungar Publishing Co., 1955.

STABILITY OF INTERPOLATING ELASTICA

Michael Golomb
Mathematics Research Center
University of Wisconsin-Madison
Madison, Wisconsin 53706

ABSTRACT

Interpolating elastica are the extremals for the functional $\int_0^{\bar{s}} \kappa^2(s)ds$, which is the integral of the square of the curvature with respect to arc length, in the family of plane curves that interpolate at (not prescribed) arc lengths $s_0 < s_1 < \dots < s_n$ a prescribed configuration of points $p_0, p_1, \dots, p_n$. If at one or both terminals the slope is prescribed, the extremal is said to be angle-constrained, otherwise free. The curvature functional represents the elastic strain energy of a thin elastic beam of indefinite length with sleeve supports anchored at $p_0, p_1, \dots, p_n$, which allow the beam to slide through without friction and to rotate freely (except at the end supports if angle-constrained). The interpolating elastica are also known as nonlinear spline curves. It is known that the infimum of the strain energy is 0 in all cases, hence cannot be attained if the points $p_0, p_1, \dots, p_n$ do not lie on a ray. On the other hand, interpolating elastica are known to exist for a variety of configurations, and this report investigates whether these extremals make the strain energy a local minimum or not (i.e., whether they are "stable" or "unstable"). Several general stability criteria are established and they are used to decide the stability of some specific elastica.

STABILITY OF INTERPOLATING ELASTICA

Michael Golomb

1. Introduction

Elastica are the plane curves with "normal representation" $s \mapsto \theta(s)$ (s denotes arc length and $\theta(s)$ the angle of inclination at s) which are solutions of the differential equation

$$\frac{1}{2}\theta'^2(s) = \lambda[\sin(\theta - \theta_1) - \alpha] \tag{1.1}$$

where λ, θ_1 and α are real constants (see, e.g. [1, Article 263]). (1.1) is the Euler equation for the variational problem

$$\delta \int_0^{\bar{s}} \theta'^2 = 0, \quad \int_0^{\bar{s}} \cos\theta = b, \quad \int_0^{\bar{s}} \sin\theta = d \tag{1.2}$$

where $\bar{s}, b, d$ are prescribed (see above reference or [2, Prop. 3.2]). The integral $\int_0^{\bar{s}} \theta'^2$ represents (with the proper choice of units) the strain energy of a thin elastic beam of uniform cross section of length $\bar{s}$, and the side conditions in (1.2) specify the relative position of the ends of the bent beam.

The elastica described by (1.1), when considered for all values of s, have infinitely many inflection points, $\theta'(s) = 0$ when $\sin(\theta(s) - \theta_1) = \alpha$, and are therefore called inflectional elastica (see [1, loc cit.]). Below we will consider only elastica for which $\alpha = 0$; geometrically speaking, these are curves for which the variation of θ between consecutive inflection points is π. We refer to them as simple elastica. All the simple elastica are obtained from a particular one by similarity transformations.

The interpolating elastica (so named by M. A. Malcolm in [3]) consist of finitely many subarcs of the simple elastica, fitted together so that a smooth curve with continuous curvature results which has jump discontinuities of the derivative of the curvature only at the "knots" $p_1, \dots, p_{n-1}$. Such an interpolating elastica E

Sponsored by the United States Army under Contract No. DAAG29-75-C-0024.

with normal representation θ is the solution of the variational problem

$$(1.3)\qquad \delta\int_{s_0}^{s_n} \theta'^2 = 0,\quad \int_{s_{i-1}}^{s_i} \cos\theta = b_i,\quad \int_{s_{i-1}}^{s_i} \sin\theta = d_i \quad (i = 1,\ldots,n)$$

where the b_i, d_i are prescribed (they are the coordinates of the vector $\overline{p_i - p_{i-1}}$), but the arc *lengths* $s_0 < s_1 < \ldots < s_n$ *of the terminals* p_0, p_n and of the knots $p_1,\ldots,p_{n-1}$ are varied (see [2, loc. cit.] or [3, Sec. 2]). If the ends p_0, p_1 are "free" then the natural boundary conditions

$$(1.3a)\qquad \theta'(0) = 0,\quad \theta'(s_n) = 0$$

are appended to (1.3). Frequently we shall be concerned with "angle-constrained" interpolating elastica; in this case we are given

$$(1.3b)\qquad \theta(0) = \alpha,\quad \theta(s_n) = \beta\ .$$

The solutions of (1.3), (1.3a) represent possible equilibrium positions (stable or not) of a thin elastic beam of indefinite length which is constrained to pass through frictionless freely rotating small sleeves anchored at the positions $p_0, p_1, \ldots, p_n$. If the sleeves at the terminals p_0, p_n are pinned then (1.3b) replaces (1.3a). The interpolating elastica are a reasonable mathematical model for the mechanical spline used by draftsmen to pass a smooth curve through the given points $p_0, p_1, \ldots, p_n$. They are also called nonlinear (interpolating) splines (see, e.g., [4]) and were referred to as <u>extremal interpolants</u> for the configuration $\{p_0, p_1, \ldots, p_n\}$ in [2]. We still will refer to them by this name in the sequel.

The solutions of (1.3) are definitely not absolute minima, *except in the trivial* case where $p_0, p_1, \ldots, p_n$ lie (in this order) along a ray (and moreover, $\alpha = \beta = 0$ in case of end conditions (1.3b)). This was first pointed out by the authors of [5]. $\int_0^{s_n} \theta'^2$ can be made arbitrarily small by using large interpolating circular loops.

The solutions are often referred to as local minima, although no proofs are given that they are indeed extrema of this kind. Only in [2, Theorem 6.1] was it proved that the nontrivial simple elastica interpolating 2 points are nonstable, i.e., they do not represent local minima of $\int_{s_0}^{s_1} \theta'^2$. It is the objective of this paper to establish, for several known extremal interpolants, whether they are local minima or not (stable or unstable).

The fact that the extremal interpolants do not represent minima nor, in general, local minima of the functional $\int \theta'^2$ is, probably, the major reason for the lack of general existence results and of good computational procedures. (For an existence proof limited to length-restricted extremals, see [6]. In [7, Theorem 3] it is proved that if there is a length-restricted extremal of "unstable length", there is also an interpolating local minimum, but no nontrivial length-restricted extremal of unstable length is presented. Existence of length-prescribed extremals and of unrestricted extremal interpolants close to a ray interpolant is proved in [2, Appendix and Theorem 7.4], where also many examples of specific interpolants are given, which were not known before. For a survey of old and new computational procedures, see [3].) In the discussion of stability (that is, whether the extremals are local minima or not) we naturally restrict ourselves to cases where existence of extremal interpolants has been proved or is postulated.

In Section 2 the variational equations for interpolating splines in normal representation are derived, without recourse to Lagrange multiplier theory, and as a preparation for the computation of the second variation. In Section 3 the second variation is used for stability criteria (Jacobi's condition): an explicitly given quadratic functional must be positive-definite, or equivalently, a nonconventional linear second-order boundary value problem must have only positive eigenvalues. In Section 4 it is proved that interpolating splines close (in a precise sense) to stable ones are stable and those close to strongly unstable ones are unstable. This result is then used to prove that splines that interpolate configurations close to a ray configuration (whose existence was proved in [2]) are stable (even in the case of free terminals). This is probably the first general existence proof for locally minimizing interpolants which are not length-restricted. In Section 5 it is proved that the extremal 2-point interpolant consisting of $n \geq 1$ complete loops of the simple elastica is unstable even if angle-constrained (in [27] the instability was proved for the free elastica). If the angle-constrained 2-point interpolants is a proper subarc of one loop of the simple elastica (hence has no inflection point) then it is stable, and any angle-constrained 2-point interpolant that contains one complete loop of the simple elastica is unstable. The proof for these last results is contained in Section 6; it is built mainly on the discovery of the eigenfunction belonging to the eigenvalue 0 for the second variational equation that goes with the one-loop angle-constrained simple elastica. By an extension of this method it is proved in Section 7 that if an angle-constrained interpolant contains an interior inflection point then it is stable if it contains neither the left nor the right "stability focus". These are points on the simple elastica which are situated symmetrically with respect to the inflection point, not far from the neighboring inflection points. If the angle-constrained 2-point interpolant with one inflection point contains both stability foci it is unstable. The general result on the stability of such 2-point interpolants is stated with the use of what we call "conjugate points". If p is a point on a simple elastica arc containing one inflection point there is a conjugate point p_* defined by a transcendental equation, and it is also given a geometric interpretation (p and p_* are on opposite sides of the inflection point; if p is a stability focus then p_* is the other stability focus). The angle-constrained elastica is stable if and only if it contains no pair of conjugate points. If the 2-point extremal interpolant is free at one end and angle-constrained at the other end, then it is stable if and only if it contains no stability focus. Section 8 contains the most important stability results. It is first proved that a necessary condition for the stability of extremal N-point interpolants is that each arc between

consecutive nodes be "proper", i.e. internal arcs do not contain a pair of conjugate points, and the terminal arcs do not contain a stability focus. Then a computable "stability function" of (N-2) real variables is defined for the extremal N-point interpolant under investigation, which has a critical value at the point that corresponds to the extremal. It is proved that the extremal is stable if and only if the critical value is a local minimum. These results are applied to decide the stability of some 3-point and 4-point extremal interpolants. In this connection it is also shown that the often repeated claim (first appearing in [5]) that a certain 4-point configuration has no interpolating elastica is false. In the last section we show that the closed extremals which interpolate the vertices of a regular n-gon (n ≠ 3) (their existence is proved in [2, Sec. 8]) are stable.

2. The Euler-Lagrange conditions for the interpolating spline in normal representation.

Let $s \mapsto s(\theta)$, $0 \leq s \leq \bar{s}$ be the normal representation of an admissable interpolant C for the configuration $\{p_0, p_1, \ldots, p_n\}$. Here s denotes the arc length along the curve C and $\theta(s)$ the angle that C makes at arc length s with a reference line. The interpolation conditions are

$$(2.1) \qquad \int_{s_{i-1}}^{s_i} \cos\theta(s)ds = b_i, \quad \int_{s_{i-1}}^{s_i} \sin\theta(s)ds = d_i, \quad i = 1,\ldots,n$$

where b_i, d_i are given numbers, and the nodes $0 = s_0 < s_1 < \ldots < s_n = \bar{s}$ are the arc lengths at which C passes through the interpolation points $p_0, p_1, \ldots, p_n$ ($s_1, \ldots, s_n$ vary with C). We assume $b_i^2 + d_i^2 > 0$, hence $p_{i-1} \neq p_i (i = 1, \ldots, n)$.

Much of the paper deals with <u>angle-constrained</u> interpolants, in which case the angles

$$(2.2) \qquad \theta(0) = \alpha, \quad \theta(\bar{s}) = \beta$$

are prescribed. If $\theta(0)$ and/or $\theta(\bar{s})$ is not prescribed the corresponding terminal of C is said to be free, and the corresponding natural end conditions for an extremal interpolant turn out to be

$$(2.3) \qquad \theta'(0) = 0, \quad \theta'(\bar{s}) = 0 .$$

The functional which is made stationary by an extremal interpolant E is the potential energy (or curvature functional)

$$(2.4) \qquad \int_0^{\bar{s}} [\theta'(s)]^2 ds .$$

The comparison functions are taken from the Sobolev space $W_{1,2} = W_{1,2}[0,S]$ of functions $\theta : [0,S] \to \mathbb{R}$, which are absolutely continuous and have derivatives θ' in $L_2[0,S]$ with norm $\{\int_0^S (\theta^2 + \theta'^2)\}^{\frac{1}{2}}$. S is a prescribed positive number large enough so that the functions in $W_{1,2}$ satisfying conditions (2.1) and (2.2) (if imposed) form a subset with nonempty interior. In this paper we do not deal with the existence of extremal interpolants, but we start with a known extremal E_0

and investigate whether it is stable or not. In this case, we may take $S = \bar{s} + \delta$ where $\bar{s}$ is the length of E_0 and δ is an arbitrary positive number.

Let $s \leftrightarrow \theta_0(s)$ be the normal representation of E_0 and $0 = s_0 < s_1 < \ldots < s_n = \bar{s}$ the interpolation nodes. For fixed real numbers $\tau_1, \ldots, \tau_n$ and fixed functions η, ξ in $W_{1,2}$, which we assume to have piecewise continuous derivatives with jumps only at $s_1, \ldots, s_{n-1}$, consider the family of comparison curves C_ε, given parametrically by

$$\theta_\varepsilon(t) = \theta_0(t) + \varepsilon\eta(t) + \varepsilon^2\xi(t), \quad 0 \leq t \leq \bar{s}$$

(2.5)
$$s_\varepsilon(t) = \sum_{j=1}^{i-1} (1 + \varepsilon\tau_j)(t_j - t_{j-1}) + (1 + \varepsilon\tau_i)(t - t_{i-1}), \quad s_{i-1} \leq t \leq s$$

where $\theta_\varepsilon(t)$, $s_\varepsilon(t)$ denote the angle of inclination and the arc length of C_ε at t. If $\varepsilon \in \mathbb{R}$ is sufficiently small then C_ε is in a prescribed neighborhood of E_0. The interpolation conditions (2.1) require

(2.6)
$$(1 + \varepsilon\tau_i) \int_{s_{i-1}}^{s_i} \cos\theta_\varepsilon(t)dt = b_i, \quad (1 + \varepsilon\tau_i) \int_{s_{i-1}}^{s_i} \sin\theta_\varepsilon(t)dt = d_i,$$
$$i = 1, \ldots, n .$$

For definiteness, we assume $d_i \neq 0 (i = 1, \ldots, n)$. Then equating terms in ε^1 in (2.6) gives

(2.7a)
$$\tau_i = -(1|d_i) \int_{s_{i-1}}^{s_i} \cos\theta_0\eta$$

(2.7b)
$$\int_{s_{i-1}}^{s_i} (b_i\cos\theta_0 + d_i\sin\theta_0)\eta = 0,$$

and equating terms in ε^2 gives

$$\int_{s_{i-1}}^{s_i} (\cos\theta_0 \eta^2 + 2\sin\theta_0 \xi) + 2b_i\tau_i^2 = 0$$

(2.7c)

$$\int_{s_{i-1}}^{s_i} (\sin\theta_0 \eta^2 - 2\cos\theta_0 \xi) + 2d_i\tau_i^2 = 0$$

$$n = 1,\dots,n \; .$$

The value of the potential energy for the curve C_ε is

(2.8)
$$U(\varepsilon) = \int_0^{\bar{s}} \left[\frac{\theta'_\varepsilon(t)}{s'_\varepsilon(t)}\right]^2 s'_\varepsilon(t)\,dt = \sum_{i=1}^{n} (1 + \varepsilon\tau_i)^{-1} \int_{s_{i-1}}^{s_i} \theta'^{\,2}_\varepsilon \; .$$

Set

(2.9)
$$u_i = \int_{s_{i-1}}^{s_i} \theta'^{\,2}_0 , \quad U_0 = \sum_{i=1}^{n} u_i \; .$$

Expand (2.8) in powers of ε, using (2.5):

(2.10)
$$U(\varepsilon) = U_0 + \varepsilon \left[2 \int_0^{\bar{s}} \theta'_0 \eta' - \sum_{i=1}^{n} \tau_i u_i\right]$$
$$+ \varepsilon^2 \left[2 \int_0^{\bar{s}} \theta'_0 \xi' - 2 \sum_{i=1}^{n} \tau_i \int_{s_{i-1}}^{s_i} \theta'_0 \eta' + \sum_{i=1}^{n} \tau_i^2 u_i + \int_0^{\bar{s}} \eta'^{\,2}\right] + O(\varepsilon^3) \; .$$

Since U_0 is a stationary value of the potential energy, we must have, using (2.7a),

$$\sum_{i=1}^{n} \int_{s_{i-1}}^{s_i} (2\theta'_0\eta' + \frac{u_i}{d_i} \cos\theta_0 \eta) = 0$$

and this must be true for every η for which (2.7b) holds and for which

(2.7d)
$$\eta(0) = 0 \quad \text{and/or} \quad \eta(\bar{s}) = 0$$

if E_0 is angle-constrained. From this one infers, by the usual arguments of the calculus of variations (carried out in detail in [2]) that θ'_0 is continuous, θ''_0 is continuous between consecutive interpolation nodes, and there exist constants $\lambda_i \in \mathbb{R}$ such that

(2.12)
$$2\theta_0''(s) - \lambda_i d_i \sin\theta_0(s) - \left(\frac{u_i}{d_i} + \lambda_i b_i\right)\cos\theta_0(s) = 0,$$
$$s_{i-1} < s < s_i, \quad i = 1,\dots,n\ .$$

Moreover, conditions (2.3) must hold for θ_0 if the terminals are free.

Integration of (2.12) gives

$$\theta_0'^2(s) + \lambda_i d_i \cos\theta_0(s) - \left(\frac{u_i}{d_i} + \lambda_i b_i\right)\sin\theta_0(s) = \delta_i$$

and another integration from s_{i-1} to s_i shows that $\delta_i = 0$. Thus,

(2.13)
$$\theta_0'^2(s) + \lambda_i^1\cos\theta_0(s) + \lambda_i^2\sin\theta_0(s) = 0,$$
$$s_{i-1} \leq s \leq s_i, \quad i = 1,\dots,n$$

where we have set

(2.14)
$$\lambda_i^1 = \lambda_i d_i, \quad \lambda_i^2 = -\lambda_i b_i - \frac{u_i}{d_i}\ .$$

To determine the multipliers λ_i^1, λ_i^2 we use the fact that θ_0 and θ_0' are continuous, hence

(2.15)
$$\left(\lambda_{i+1}^1 - \lambda_i^1\right)\cos\theta_0(s_i) + \left(\lambda_{i+1}^2 - \lambda_i^2\right)\sin\theta_0(s_i) = 0,$$
$$i = 1,\dots,n-1\ .$$

Conditions (2.15) together with the interpolation conditions (2.1) and end conditions $\theta_0(0) = \alpha(\text{or } \theta_0'(0) = 0)$, $\theta_0(\bar{s}) = \beta(\text{or } \theta_0'(\bar{s}) = 0)$, are $3n + 1$ independent conditions for the $3n + 1$ unknowns $\lambda_i^1, \lambda_i^2, s_i (i = 1,\dots,n)$ and $\theta_0(0)$, which together with the differential equation (2.13) determine the interpolating elastica θ_0. There may be many solutions of these equations, as shown in [2], but the distinct solutions are isolated.

The assumption $d_i \neq 0 (i = 1,\dots,n)$ was made only to avoid case splitting. The obtained result remains true as long as $b_i^2 + d_i^2 > 0$ for $i = 1,\dots,n$.

3. Stability Criteria.

We now look at the quadratic terms in the expansion (2.10) for the potential energy:

(3.1)
$$\int_0^{\bar s} \eta'^2 - 2\sum_{i=1}^{n} \tau_i \int_{s_{i-1}}^{s_i} \theta_0'\eta' + \sum_{i=1}^{n} \tau_i^2 u_i + 2\int_0^{\bar s} \theta_0'\xi',$$

$$\tau_i = -\frac{1}{d_i}\int_{s_{i-1}}^{s_i} (\cos\theta_0)\eta = \frac{1}{b_i}\int_{s_{i-1}}^{s_i} (\sin\theta_0)\eta, \quad u_i = \int_{s_{i-1}}^{s_i} \theta_0'^2 \ . \ .$$

Using (2.7c), (2.12), (2.13) and (2.14), we can eliminate ξ in (3.1):

(3.2)
$$2\int_{s_{i-1}}^{s_i} \theta_0'\xi'' - 2\theta_0'\Big|_{s_{i-1}}^{s_i} = -2\int_{s_{i-1}}^{s_i} \theta_0''\xi$$

$$= -\lambda_i d_i \int_{s_{i-1}}^{s_i} (\sin\theta_0)\xi - \left(\frac{u_i}{d_i} + \lambda_i b_i\right)\int_{s_{i-1}}^{s_i} (\cos\theta_0)\xi$$

$$= \lambda_i d_i \left[\frac{1}{2}\int_{s_{i-1}}^{s_i} (\cos\theta_0)\eta^2 + b_i\tau_i^2\right] - \left(\frac{u_i}{d_i} + \lambda_i b_i\right)\left[\frac{1}{2}\int_{s_{i-1}}^{s_i} (\sin\theta_0)\eta^2 + d_i\tau_i^2\right]$$

$$= -\frac{1}{2}\int_{s_{i-1}}^{s_i} \theta_0'^2\eta^2 - \tau_i^2 u_i \ .$$

Thus, since $(\theta_0'\xi)(s_i - 0) = (\theta_0'\xi)(s_i + 0)$ and $(\theta_0'\xi)(0) = (\theta_0'\xi)(\bar s) = 0$:

$$2\int_0^{\bar s} \theta_0'\xi' = -\frac{1}{2}\int_0^{\bar s} \theta_0'^2\eta^2 - \sum_{i=1}^{n} \tau_i^2 u_i$$

and (3.1) becomes

(3.3)
$$\int_0^{\bar s} (\eta'^2 - 2\sum_{i=1}^{n} \tau_i \int_{s_{i-1}}^{s_i} \theta_0'\eta') - \frac{1}{2}\int_0^{\bar s} \theta_0'^2\eta^2 \ .$$

We introduce the subspace $V(\theta_0)$ of $W_{1,2}[0,\bar s]$:

$$(3.4)\quad V_0(\theta_0) = \{\eta \in W_{1,2}[0,\bar{s}] : \int_{s_{i-1}}^{s_i} (b_i\cos\theta_0 + d_i\sin\theta_0)\eta = 0 \text{ for } i = 1,\dots,n;$$

$$\eta(0) = 0 \text{ and/or } \eta(\bar{s}) = 0 \text{ if } E_0 \text{ is angle-constrained at the corresponding terminal}\}.$$

and the quadratic form $Q(\theta_0,\cdot)$ with domain V_0:

$$(3.5)\qquad Q(\theta_0,\eta) = \int_0^{\bar{s}} (\eta'^2 - \tfrac{1}{2}\theta_0'^2\eta^2) + 2\sum_{i=1}^{n} d_i^{-1}\int_{s_{i-1}}^{s_i} (\cos\theta_0)\eta \int_{s_{i-1}}^{s_i} \theta_0'\eta' .$$

It is understood that the factor $d_i^{-1}\int_{s_{i-1}}^{s_i} (\cos\theta_0)\eta$ is replaced by $-b_i^{-1}\int_{s_{i-1}}^{s_i} (\sin\theta_0)\eta$ if $d_i = 0$. If $Q(\theta_0,\eta) \le 0$ for some $\eta \neq 0$ then the stationary value U_0 is not a strict local minimum of the potential energy, i.e. the extremal interpolant E_0 is not stable. If $Q(\theta_0,\eta) > 0$ for each $\eta \neq 0$ then the potential energy is larger than U_0 for every admissable interpolant $\theta \neq \theta_0$ in some $W_{1,2}[0,S]$ neighborhood of θ_0 (not only for those of the form (2.5)), hence U_0 is a strict local minimum and E_0 is a stable extremal. We have obtained

Proposition 3.1. The possibly angle-constrained extremal interpolant E_0 with the normal representation $s \mapsto \theta_0(s)$, interpolation nodes $0 = s_0 < \dots < s_n = \bar{s}$, and interpolation data $\int_{s_{i-1}}^{s_i} \cos\theta_0 = b_i$, $\int_{s_{i-1}}^{s_i} \sin\theta_0 = d_i$, is stable if and only if the quadratic form (3.5) with domain (3.4) is positive definite, i.e. $Q(\theta_0,\eta) > 0$ for every $\eta \neq 0$.

Set now

$$(3.6)\qquad Q_* = \inf_{\eta\in V_0,\ \int_0^{\bar{s}}\eta^2=1} Q(\theta_0,\eta) .$$

Clearly $Q_* > -\infty$. Also $\int_0^{\bar{s}} \eta'^2$ is bounded for $\eta \in V(\theta_0)$, $\int_0^{\bar{s}} \eta^2 = 1$, $Q(\theta_0,\eta) \le Q_* + 1$.

By familiar arguments it follows that the continuous form Q attains the value Q_* for some $\eta_* \in V_0(\theta_0)$, $\int_0^{\bar{s}} \eta_*^2 = 1$. The Euler equation for η_* is:

$$\eta_*''(s) + \frac{1}{2}\theta_0'^2(s)\eta_*(s) - d_i^{-1}\left(\int_{s_{i-1}}^{s_i} \theta_0'\eta_*'\right)\cos\theta_0(s) + d_i^{-1}\left(\int_{s_{i-1}}^{s_i} \cos\theta_0\eta_*\right)\theta_0''(s)$$

$$+ \rho_i[b_i\cos\theta_0(s) + d_i\sin\theta_0(s)] + \mu_*\eta_*(s) = 0,$$

$$s_{i-1} \le s \le s_i, \quad i = 1,\ldots,n; \quad \eta_*' \text{ continuous} .$$

The multipliers $\rho_i \in \mathbb{R}$ result from the side conditions $\int_{s_{i-1}}^{s_i} (b_i\cos\theta_0 + d_i\sin\theta_0)\eta = 0$, and $\mu_* \in \mathbb{R}$ from the condition $\int_0^{\bar{s}} \eta^2 = 1$. It should be understood that $d_i^{-1}\cos\theta_0$ in the two integral terms of (3.7a) is replaced by $-b_i^{-1}\sin\theta_0$ if $d_i = 0$. (3.7a) is supplemented by the conditions of (3.4)

(3.7b) $$\int_{s_{i-1}}^{s_i} (b_i\cos\theta_0 + d_i\sin\theta_0)\eta_* = 0, \quad i = 1,\ldots,n,$$

and

(3.7c) $$\eta(0) = 0 \text{ or } \eta_*'(0) = 0 \quad \text{and} \quad \eta_*(\bar{s}) = 0 \text{ or } \eta_*'(\bar{s}) = 0$$

depending on whether θ_0 is angle-constrained or free. Besides we have the conditions

(3.7d) $$\theta_0'(s_i)\left[d_i^{-1}\int_{s_{i-1}}^{s_i} \cos\theta_*\eta_* - d_{i+1}^{-1}\int_{s_i}^{s_{i+1}} \cos\theta_0\eta_*\right] = 0, \quad i = 1,\ldots,n-1$$

resulting from the fact that the coefficient of η' in (3.5) is discontinuous. If $\theta_0'(s_i) \neq 0$ for the internal nodes $s_1,\ldots,s_{n-1}$ then (3.7d) combined with (3.7c) requires:

(3.8) $$d_i^{-1}\int_{s_{i-1}}^{s_i} \cos\theta_0\eta_* = -b_i^{-1}\int_{s_{i-1}}^{s_i} \sin\theta_0\eta_* = \text{constant for } i = 1,\ldots,n .$$

The multipliers ρ_i can be eliminated from (3.7a). We integrate (3.7a) over the interval (s_{i-1}, s_i) and obtain:

$$\rho_i(b_i^2 + d_i^2) = \eta_*'(s_{i-1}) - \eta_*'(s_i) - \frac{1}{2}\int_{s_{i-1}}^{s_i} \theta_0'^2 \eta_* + b_i d_i^{-1} \int_{s_{i-1}}^{s_i} \theta_0' \eta_*'$$

(3.9)

$$+ d_i^{-1}(\theta_0'(s_{i-1}) - \theta_0'(s_i)) \int_{s_{i-1}}^{s_i} \cos\theta_0 \eta_* - \mu_* \int_{s_{i-1}}^{s_i} \eta_* .$$

With (3.9) substituted in (3.7a), we obtain the equation

$$\eta_*''(s) + \frac{1}{2}\theta_0'^2 \eta_*(s) + \beta_i(\eta_*)\cos\theta_0(s) + \delta_i(\eta_*)\sin\theta_0(s) + \mu_*\eta_*(s) = 0,$$

(3.10)

$$s_{i-1} \le s \le s_i, \quad i + 1, \dots, n$$

where the β_i and δ_i are well-defined linear functionals, depending only on θ_0. (3.10) together with (3.7b,c,d) is a nonconventional linear boundary-value problem for η_*, μ_* being the eigenvalue. Introduce the linear operator R with domain $\mathcal{D}(R)$ of functions $\eta : [0,\bar{s}] \to \mathbb{R}$, with η' continuous on $[0,\bar{s}]$, η'' continuous on each $[s_{i-1}, s_i]$ and η satisfying conditions (3.7b,c,d), defined by:

$$(R\eta)(s) = -\eta''(s) - \frac{1}{2}\theta_0'^2(s)\eta(s) - \beta_i(\eta)\cos\theta_0(s) - \delta_i(\eta)\sin\theta_0(s),$$

(3.11)

$$s_{i-1} < s < s_i, \quad i = 1, \dots, n .$$

Then the above eigenvalue problem may be stated as:

(3.12) $$R\eta = \mu\eta .$$

A simple calculation shows that if $\int_0^{\bar{s}} \eta^2 = 1$ then

(3.13) $$\mu = \int_0^{\bar{s}} \eta R\eta = Q(\theta_0, \eta) .$$

Therefore, $\mu_* = Q(\theta_0, \eta_*)$ is the smallest eigenvalue of R.

We conclude that the form Q is positive definite if and only if $\mu_* > 0$, *or equivalently,* all the eigenvalues of R are positive. We have obtained

<u>Proposition 3.2.</u> The extremal interpolant E_0 is stable if and only if the operator R defined above has only positive eigenvalues.

The following proposition provides a useful sufficient condition for instability of interpolating elastica.

Proposition 3.3. Suppose E_0 with normal representation $s \mapsto \theta_0(s)$ $(s_1 \leq s \leq s_n)$ is an angle-constrained extremal interpolant for some configuration $\{p_1, p_2, \ldots, p_n\}$. Suppose E is another extremal interpolant (angle-constrained or free), with normal representation $s \mapsto \theta(s)$ $(s_0 \leq s \leq s_{n+1},\ s_0 \leq s_1,\ s_{n+1} \geq s_n)$, where θ is an extension of θ_0 with no additional knot; thus, θ'' is continuous at $s_1 (s_n)$ if $s_0 < s_1$ $(s_{n+1} > s_n)$. Then E is also unstable.

Proof. The extremal E, which interpolates the configuration $\{p_0, p_2, \ldots, p_{n-1}, p_{n+1}\}$ can also be considered as an extremal $\bar{E}$ which interpolates the configuration with $p_1 (p_n)$ inserted between p_0 and p_2 $(p_{n-1}$ and $p_{n+1})$ if $s_0 < s_1.$ $(s_{n+1} > s_n)$. Let $\bar{\theta}$ denote θ in this identification. Since E_0 is unstable there exists, by Proposition 3.1, $\eta_0 \in V_0(\theta_0)$ such that $Q(\theta_0, \eta_0) \leq 0$. In particular, $\eta_0(s_1) = \eta_0(s_n) = 0$. Let $\bar{\eta}$ be defined as an extension of η_0:

$$\begin{aligned} \bar{\eta}(s) &= \eta_0(s), \quad s_1 \leq s \leq s_n \\ &= 0, \quad s_0 \leq s < s_1 \text{ and } s_n < s \leq s_{n+1}. \end{aligned} \tag{3.14}$$

It is easily checked that $\bar{\eta} \in V_0(\bar{\theta})$ and $Q(\bar{\theta}, \bar{\eta}) = Q(\theta_0, \eta_0) \leq 0$. It follows, again by Proposition 3.1, that $\bar{E}$ is unstable. Since E is obtained from $\bar{E}$ by the removal of constraints, E is unstable.

Let E_0 of Proposition 3.3 be angle-constrained at one terminal only, say at p_n. For this case we have the

Corollary. If the unstable extremal of Proposition 3.3 is angle-constrained only at p_n and θ is an extension of θ_0 to $s_0 \leq s \leq s_{n+1}$ with θ'' continuous at s_n then E is unstable.

The proof of this is an obvious modification of that for Proposition 3.3.

Another useful sufficient condition is expressed in the following

Proposition 3.4. Suppose E is an interpolating elastica angle-constrained at none, one, or both terminals) and E_i is a subarc between consecutive nodes of E. If E_i, considered as a 2-point extremal interpolant which is angle-constrained at the terminals which are internal nodes of E, is unstable then E is.

<u>Proof.</u> Suppose $s \mapsto \theta(s)$ $(0 \leq s \leq \bar{s})$ is the normal representation of E and $s \mapsto \theta_i(s)$ $(s_{i-1} \leq s \leq s_i)$ is the restriction of θ which represents E_i. Since E_i is unstable there exists $\eta_i \in V_0(\theta_i)$, $\eta_i \neq 0$, such that $Q(\theta_i,\eta_i) \leq 0$. In particular, $\eta_i(s_{i-1}) = 0$ and/or $\eta_i(s_i) = 0$ if $i \geq 2$ and/or $i \leq n - 1$. The extension η_i with value 0 on $[0,s_{i-1})$ (if $i \geq 2$) and on $(s_i,\bar{s}]$ (if $i \leq n - 1$) is continuous, and clearly $\eta \in V_0(\theta)$, $Q(\theta,\eta) = Q(\theta_i,\eta_i) \leq 0$. Hence E is unstable.

4. Extremals close to stable ones.

If E_0 is an extremal interpolant of some configuration $\{p_0, p_1, \ldots, p_n\}$, $s \mapsto \theta_0(s)$, $0 \le s \le \bar{s}$, is its normal representation, and $0 = s_0 < s_1 < \ldots < s_n = \bar{s}$ are its interpolation nodes, then $t = \theta_0(\bar{s}t) = \tilde{\theta}_0(t)$, $0 \le t \le 1$, is the normal representation of an extr^mal interpolant $\tilde{E}_0$ for the configuration $\{\bar{s}^{-1}p_0, \bar{s}^{-1}p_1, \ldots, \bar{s}^{-1}p_n\}$, with interpolation nodes $0 = t_0 < t_1 < \ldots < t_n = 1$, $t_i = s_i|\bar{s}$. If the terminals of $\tilde{E}_0$ are free or angle-constrained, so are those of E_0. Clearly, $\tilde{E}_0$ is stable if and only if E_0 is. In the following we will often use the standardized normal representation of elastica.

Let F_1 denote the metric space of functions $\theta\colon [0,1] \to \mathbb{R}$, for which there are real numbers $\alpha = \alpha(\theta)$, $\beta = \beta(\theta)$ such that the equations

$$\begin{aligned} &\tfrac{1}{2}\theta'^2(t) = \alpha\sin\theta(t) - \beta\cos\theta(t) \\ (4.1)\qquad &\theta'(t) = \alpha\cos\theta(t) + \beta\sin\theta(t), \quad 0 \le t \le 1, \end{aligned}$$

hold with the distance functional $d_1(\theta_1, \theta_2) = \max_{0 \le t \le 1}|\theta_1(t) - \theta_2(t)|$. For the proof of the proposition below we will use the following

Lemma. The functionals α, β from F_1 to $\mathbb{R}$ are continuous.

Proof. First, $\alpha(\theta), \beta(\theta)$ are uniquely defined, for if (4.1) and also $\frac{1}{2}\theta'^2 = \alpha_1\sin\theta - \beta_1\cos\theta$, $\theta'' = \alpha_1\cos\theta + \beta_1\sin\theta$ hold, then $0 = (\alpha - \alpha_1)\sin\theta(t) - (\beta - \beta_1)\cos\theta(t) = (\alpha - \alpha_1)\cos\theta(t) + (\beta - \beta_1)\sin\theta(t)$, hence $\alpha = \alpha_1$ and $\beta = \beta_1$. Clearly, (4.1) is equivalent to

$$(4.2)\qquad \theta(t) - (1 - t)\theta(0) - t\theta(1) = \int_0^1 g(t,\tau)\{\alpha\cos\theta(\tau) + \beta\sin\theta(\tau)\}d\tau,$$

where

$$\begin{aligned} g(t,\tau) &= (t - 1)\tau, \quad 0 \le \tau \le t \\ &= (\tau - 1)t, \quad t \le \tau \le 1 . \end{aligned}$$

The uniqueness of α, β in (4.2) implies that the continuous functions $x = \int_0^1 g(\cdot,\tau)\cos\theta(\tau)d\tau$, $y = \int_0^1 g(\cdot,\tau)\sin\theta(\tau)d\tau$ are linearly independent.

Therefore, the Gramian $\int x^2 \int y^2 - (\int xy)^2$ is $\neq 0$. Thus, if (4.2) is dot-multiplied by x and y respectively, two independent linear scalar equations for α, β are obtained, whose solution demonstrates the assertion of the Lemma.

We also observe that the functionals $\alpha(\theta), \beta(\theta)$ are uniquely determined by the restriction of θ to any subinterval of $[0,1]$.

Now let F_n denote the class of interpolating elastica E with normal representation $t \mapsto \theta(t)$, $0 \leq t \leq 1$, all satisfying the same type of end conditions (free or angle constraints) and having $(n + 1)$ interpolation nodes $0 = t_0 < t_1 < \ldots < t_n = 1$, where $t_i = t_i(\theta)$ for $i = 1, \ldots, n - 1$. We also assume that no two consecutive interpolation points of E coincide. F_n is made into a metric space by use of the distance functional

$$d(\theta_1, \theta_2) = \max_{i=1,\ldots,n-1} |t_i(\theta_1) - t_i(\theta_2)| + \max_{\underline{0 \leq t \leq 1}} |\theta_1(t) - \theta_2(t)| \ . \tag{4.4}$$

We now prove

<u>Proposition 4.1.</u> Suppose E_0 is an interpolating elastica with normal representation $\theta_0 \in F_n$, which is stable. Then there exists $\delta > 0$ such that every interpolating elastica E with normal representation $\theta \in F_n$ for which $d(\theta_0, \theta) < \delta$ is also stable.

<u>Proof.</u> For every $\theta \in F_n$ we have by (2.12)

$$\begin{aligned} &\tfrac{1}{2}\theta'^2(t) = \alpha_i(\theta)\sin\theta(t) - \beta_i(\theta)\cos\theta(t), \\ (4.5) \qquad &\theta''(t) = \alpha_i(\theta)\cos\theta(t) + \beta_i(\theta)\sin\theta(t), \\ &t_{i-1}(\theta) \leq t \leq t_i(\theta), \quad i = 1, \ldots, n \ . \end{aligned}$$

We first take $\delta_1 > 0$ so that $d(\theta, \theta_0) < \delta_1$ implies

$$\big(t_{i-1}(\theta_0), t_i(\theta_0)\big) \cap \big(t_{i-1}(\theta), t_i(\theta)\big) \neq \phi \quad \text{for } i = 1, \ldots, n \ . \tag{4.6}$$

It then follows from the above Lemma that we can find $\delta_2 > 0$, $\delta_2 < \delta_1$ so that for $d(\theta, \theta_0) < \delta_2$:

$$|\alpha_i(\theta) - \alpha_i(\theta_0)| < 1, \quad |\beta_i(\theta) - \beta_i(\theta_0)| < 1, \quad i = 1, \ldots, n,$$

hence by (4.5)

$$\max_{0\le t\le 1} |\theta''(t)| \le M$$

for some M. Thus, the family $\{\theta' : \theta \in F_n, d(\theta,\theta_0) < \delta_2\}$ is equicontinuous:

$$(4.7) \qquad |\theta'(t') - \theta'(t'')| \le M|t' - t''| \ .$$

Suppose $\varepsilon > 0$ is given. Using the Lemma again and Equations (4.5), we can choose $\delta_3 > 0$, $\delta_3 \le \delta_2$ so that $d(\theta,\theta_0) < \delta_3$ implies $|\alpha_i(\theta) - \alpha_i(\theta_0)| + |\beta_i(\theta) - \beta_i(\theta_0)|$ is so small for $i = 1,\dots,n$ that (4.5) yields

$$(4.8) \qquad |\theta'(t) - \theta_0'(t)| < \frac{\varepsilon}{3} \quad \text{for} \quad t \in [t_{i-1}(\theta_0), t_i(\theta_0)] \cap [t_{i-1}(\theta), t_i(\theta)],$$
$$i = 1,\dots,n \ .$$

Let the overlapping of the two intervals in (4.8) occur so that $t_{i-1}(\theta_0) \le t_{i-1}(\theta) < t_i(\theta_0) \le t_i(\theta)$. If $\delta_4 > 0$, $\delta_4 \le \delta_3$ is such that $M|t_{i-1}(\theta) - t_{i-1}(\theta_0)| < \varepsilon/3$ for $d(\theta,\theta_0) < \delta_4$ then by (4.7) and (4.8), for $t_{i-1}(\theta_0) \le t \le t_{i-1}(\theta)$:

$$|\theta'(t) - \theta_0'(t)| \le |\theta_0'(t) - \theta_0'(t_{i-1}(\theta_0))| + |\theta_0'(t_{i-1}(\theta_0)) - \theta'(t_{i-1}(\theta_0))|$$
$$+ |\theta'(t_{i-1}(\theta_0)) - \theta'(t)| < \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3}$$

and the same result is obtained for $t_i(\theta_0) \le t \le t_i(\theta)$. Altogether one finds that for $d(\theta,\theta_0) < \delta_4$:

$$(4.9) \qquad |\theta'(t) - \theta_0'(t)| < \varepsilon, \quad 0 \le t \le 1 \ .$$

For $\theta \in F_n$, $\eta \in W_{1,2}[0,1]$ define (compare (3.5)):

$$(4.10) \qquad Q(\theta,\eta) = \int_0^1 (\eta'^2 - \tfrac{1}{2}\theta'^2\eta^2) + 2\sum_{i=1}^{n} \left[\int_{t_{i-1}}^{t_i} (\cos\theta)\eta \Big/ \int_{t_{i-1}}^{t_i} \sin\theta\right] \int_{t_{i-1}}^{t_i} \theta\eta',$$

where t_i stands for $t_i(\theta)$. In (4.10) it is assumed that $\int_{t_{i-1}}^{t_i} \sin\theta \ne 0$; if $\int_{t_{i-1}}^{t_i} \sin\theta = 0$ then the term in brackets is to be replaced by

$- \left[\int_{t_{i-1}}^{t_i} (\sin\theta)\eta \Big/ \int_{t_{i-1}}^{t_i} \cos\theta \right]$. It follows from (4.9) that one can find, for a given bounded set $B \subset W_{1,2}[0,1]$, $\delta_5 > 0$, $\delta_5 \leq \delta_4$, such that

$$|Q(\theta,\eta) - Q(\theta_0,\eta)| < 2\varepsilon \tag{4.11}$$

for all $\eta \in B$ and $\theta \in F_n$, $d(\theta,\theta_0) < \delta_5$.

For $\theta \in F_n$ we also define the subspace $V_0(\theta)$ of $W_{1,2}[0,1]$ (see (3.4)):

$$V_0(\theta) = \{\eta \in W_{1,2}[0,1] : \int_{t_{i-1}}^{t_i} dt \int_{t_{i-1}}^{t_i} d\tau\eta(t)\cos[\theta(t) - \theta(\tau)] = 0,$$

$$i = 1,\ldots,n; \text{ and } \eta(0) = 0 \text{ and/or } \eta(1) = 0 \text{ if the elements of } F_n \text{ are angle-constrained at the corresponding terminal}\}. \tag{4.12}$$

For the given bounded set $B \subset W_{1,2}[0,1]$ (B is then totally bounded in $L_2[0,1]$) one can choose $\delta_6 > 0$, $\delta_6 \leq \delta_5$, so that the L_2 - distance of the sets $V_0(\theta) \cap B$, $V_0(\theta_0) \cap B$ is arbitrary small if $d(\theta,\theta_0) < \delta_6$. From this, together with (4.11), one concludes that $\delta > 0$ can be found such that $d(\theta,\theta_0) < \delta$ implies

$$\inf_{\eta\in V_0(\theta_0),\, \int_0^1 \eta^2=1} Q(\theta_0,\eta) \quad - \quad \inf_{\eta\in V_0(\theta),\, \int_0^1 \eta^2=1} Q(\theta,\eta) \quad < \quad \inf_{\eta\in V_0(\theta_0),\, \eta^2=1} Q(\theta_0,\eta), \tag{4.13}$$

hence that, by Proposition 3.1, E is stable.

Proposition 4.2. If θ_0 in Proposition 3.1 is strongly unstable (i.e. $\inf_{\eta\in V(\theta_0),\, \int\eta^2=1} Q(\theta_0,\eta) < 0$), then there is $\delta > 0$ such that the elastica E for which $d(\theta_0,\theta) < \delta$ are also unstable.

We apply Proposition 4.1 to extremals which interpolate configurations close to the ray configuration. Suppose E_0 is the extremal interpolant with normal representation $\theta_0(t) = 0$, $0 \leq t \leq 1$, which interpolates the ray configuration $\{p_0^0, p_1^0, \ldots, p_n^0\}$, where $p_i^0 = (t_i^0, 0)$, $0 = t_0^0 < t_1^0 < \ldots < t_n^0 = 1$, and has free terminals. It was proved in [2, Theorem 7.4] that, given $\varepsilon > 0$, there exists $\delta > 0$ such that for every configuration $\{p_0, p_1, \ldots, p_n\}$ with $|p_i - p_i^0| < \delta$ there is a unique extremal interpolant E_ε with free ends and normal representation θ_ε for which $d(\theta_0, \theta_\varepsilon) < \varepsilon$. Now θ_0 is stable. In fact, $Q(\theta_0, \eta) = \int_0^1 \eta'^2$ and $V(\theta_0) = \{\eta : \int_{t_{i-1}}^{t_i} \eta = 0,\ i = 1, \ldots, n\}$. In particular, we must have $\int_0^1 \eta = 0$ for $\eta \in V(\theta_0)$, and it follows that $Q(\theta_0, \eta) \geq 4\pi^2 \int_0^1 \eta^2$. Thus we have obtained

Proposition 4.3. For every configuration sufficiently close to the ray configuration there exists a unique stable extremal interpolant with free terminals that is close to the trivial interpolant.

Of course, this proposition holds, a fortiori, for extremal interpolants with angle constraints.

5. Instability of the 2-point interpolants E_n, E_n^*.

If the configuration to be interpolated consists of two points p_0, p_1 then the elastica E_0 has normal representation $\theta_0 \in C_2[0,1]$, satisfying the equations (see (2.12), (2.13)):

$$\tfrac{1}{2}\theta_0'^2 - \lambda^1\sin\theta_0 + \lambda^2\cos\theta_0 = 0, \quad \theta_0'' - \lambda^1\cos\theta_0 - \lambda^2\sin\theta_0 = 0 .$$

In the sequel we will arrange it so that $\theta_0' = 0$ when $\theta_0 = 0$ or π; then these equations become

$$\tfrac{1}{2}\theta_0'^2 = \lambda_0\sin\theta_0, \quad \theta_0'' = \lambda_0\cos\theta_0 \tag{5.1}$$

for some $\lambda_0 \in \mathbb{R}$. If $p_0 = (0,0)$, $p_1 = (0,d)$, $d > 0$, then the interpolation conditions are

$$\int_0^1 \cos\theta_0 = 0, \quad \int_0^1 \sin\theta_0 = d .$$

If $\theta_0(0) = \alpha$, $\theta_0(1) = \beta$, $0 \leq \alpha \leq \pi$, $0 \leq \beta \leq \pi$, then by (5.1):

$$(2\lambda_0)^{1/2} = \left|\int_\alpha^\beta \sin^{-1/2}u\,du\right|, \quad (2\lambda_0)^{1/2}d = \left|\int_\alpha^\beta \sin^{1/2}u\,du\right| . \tag{5.2}$$

However, these formulas for λ_0 and d are correct only if $\theta_0'(t) \neq 0$ for $0 < t < 1$ (i.e. E_0 has no internal inflection point); otherwise they must be modified, as will be done below.

The quadratic form (3.5) becomes

$$Q(\theta_0,\eta) = \int_0^1 (\eta'^2 - \lambda_0\sin\theta_0\eta^2) - (2\lambda_0/d)\left(\int_0^1 \cos\theta_0\eta\right)^2 \tag{5.3}$$

and it is to be minimized on the space (see (3.4)):

$$V_0(\theta_0) = \{\eta \in W_{1,2}[0,1] : \int_0^1 \eta\sin\theta_0 = 0; \tag{5.4}$$

$$\eta(0) = 0 \text{ and/or } \eta(1) = 0 \text{ if } E_0 \text{ is angle-constrained}\} .$$

a. We first investigate the stability of the extremal $E_n (n \geq 1)$ with free terminals which has $(n - 1)$ internal and 2 terminal inflection points. E_n consists of n arcs, congruent to E_1, which is the basic nontrivial 2-point extremal interpolant (see [2, Sec. 5]). If θ_n is the normal representation of E_n and we choose $\theta_n(0) = 0$ then $\theta_n(t)$ varies from 0 to π to 0 to $\pi\ldots$ to $\frac{1}{2}[1 - (-1)^n]\pi$ as t varies from 0 to $1/n$ to $2/n$ to ... to n/n. The points $k/n (k = 1,\ldots,n - 1)$ are the internal inflection points. The total variation of θ_n is $Va(\theta_n) = n\pi$. We have

$$(5.5) \qquad \frac{1}{2}\theta_n'^2(t) = \lambda_n \sin\theta_n(t)$$

$$\theta_n(1/n + t) = \pi - \theta_n(t), \quad \theta_n(2/n + t) = \theta_n(t)$$

$$\theta_n(0) = 0 .$$

Formulas (5.2) are now replaced by

$$(5.6) \qquad (2\lambda_n)^{1/2} = n\int_0^\pi \sin^{-1/2}u\, du, \quad (2\lambda_n)^{1/2}d = n\int_0^\pi \sin^{1/2}u\, du .$$

We choose

$$\eta = \theta_n - d^{-1}\int_0^1 \theta_n \sin\theta_n .$$

Then $\int_0^1 \eta\sin\theta_n = 0$, hence $\eta \in V_0(\theta_n)$. Since $\int_0^1 \cos\theta_n = 0$, we have $\int_0^1 \eta\cos\theta_n = \int_0^1 \theta_n\cos\theta_n$, and (5.3) becomes

$$(5.8) \qquad Q(\theta_n,\eta) = 2\lambda_n d - \lambda_n\int_0^1 \theta_n^2\sin\theta_n + (\lambda_n/d)\left(\int_0^1 \theta_n\sin\theta_n\right)^2 - (2\lambda_n/d)\left(\int_0^1 \theta_n\cos\theta_n\right)^2 .$$

We use

$$\left(\int_0^1 \theta_n\sin\theta_n\right)^2 \le \int_0^1 \theta_n^2\sin\theta_n \int_0^1 \sin\theta_n = d\int_0^1 \theta_n^2\sin\theta_n$$

and find

$$(5.9) \qquad Q(\theta_n,\eta) \le (2\lambda_n/d)\left[d^2 - \left(\int_0^1 \theta_n\cos\theta_n\right)^2\right] .$$

To evaluate the integral term in (5.9) we first assume n even. Then

$$(5.10)\quad \int_0^1 \theta_n \cos\theta_n = \frac{n}{2}\left[\int_0^{1/n} \theta_n \cos\theta_n - \int_0^{1/n} (\pi - \theta_n)\cos\theta_n\right]$$

$$= -(n/2\lambda_n)2\int_0^{1/n} \theta_n'^2 = -2d\ .$$

We find the same result for n odd. (5.9), (5.10) show $Q(\theta_n,\eta) < 0$. Thus, we have proved that E_n is unstable. This was also proved in [2], but by a different method.

b. We now show that the above extremal is, for $n \geq 2$, also unstable if angle-constrained at both ends. Let this extremal be denoted as E_n^*. If θ_n^* is its normal representation then θ_n^* minimizes $\int_0^1 \theta'^2$ among the functions that satisfy the interpolation conditions $\int_0^1 \cos\theta = 0$, $\int_0^1 \sin\theta = d$ and the end conditions $\theta(0) = 0$, $\theta(1) = \frac{1}{2}[1 - (-1)^n]\pi$. E_n^* coincides with E_n of paragraph a., hence $\theta_n^* = \theta_n$. $\eta \in V(\theta_n^*)$ now requires $\eta(0) = \eta(1) = 0$ besides $\int_0^1 \eta\sin\theta_n = 0$. We choose

$$(5.11)\quad \eta_*(t) = \theta_n'(t),\quad 0 \leq t \leq 2/n$$

$$= 0,\quad 2/n \leq t \leq 1\ .$$

Then, clearly, $\eta_* \in V_0(\theta_n^*)$, and also $\int_0^1 \eta_*\cos\theta_n = 0$. Thus (5.3) becomes

$$(5.12)\quad Q(\theta_n^*,\eta_*) = \int_0^{2/n} (\eta_*'^2 - \lambda_n \sin\theta_n \eta_*^2)$$

$$= \int_0^{2/n} (\lambda_n^2\cos^2\theta_n - 2\lambda_n^2\sin^2\theta_n) = \lambda_n^2[2/n - 3\int_0^{2/n} \sin^2\theta_n].$$

But, using (5.5), (5.6) and integration by parts, we find

$$(5.13)\quad \int_0^{2/n} \sin^2\theta_n = (2/\sqrt{2\lambda_n})\int_0^{\pi} \sin^{3/2}u\,du = (2/3\sqrt{2\lambda_n})\int_0^{\pi} \sin^{-1/2}u\,du = 2/3n\ .$$

Thus, $Q(\eta_*) = 0$, and this proves instability of the extremal E_n^*, $n \geq 2$.

In the next section it will be proved that E_1^* is also unstable.

6. Two-point angle-constrained interpolants with no inflection point.

In this section we prove that 2-point angle-constrained interpolants are stable if they have no inflection point, and are unstable if they have at least 2 inflection points.

Proposition 6.1. A 2-poirt angle-constrained extremal interpolant E with no inflection point is stable.

Proof. If E has no inflection point then E is a proper subarc of the basic 2-point extremal E_1 (see Sec. 5). Clearly E is contained in another proper subarc $\check{E}$ of E_1 which has an axis of symmetry. By Proposition 3.3 it suffices to prove that the angle-constrained extremal $\check{E}$ is stable. Let $t \mapsto \check{\theta}(t)$ $(0 \le t \le 1)$ be the normal representation of $\check{E}$ and $(0,0)$, $(0,d)$, $(d > 0)$ the coordinates of the terminals, with $\theta = 0$ along the positive x axis. Then we have the following equations for $\check{\theta}$:

$$\begin{aligned}
&\tfrac{1}{2}\check{\theta}'^2(t) = \lambda\sin\check{\theta}(t), \quad 0 \le t \le 1 \\
&\check{\theta}(t) = \pi - \check{\theta}(1-t) \\
(6.1)\qquad &\check{\theta}(0) = \alpha, \quad 0 < \alpha < \pi/2 \\
&d = \int_0^1 \sin\check{\theta} = 2(2\lambda)^{-1/2} \int_\alpha^{\pi/2} \sin^{1/2}u\, du \\
&(2\lambda)^{1/2} = 2\int_\alpha^{\pi/2} \sin^{-1/2}u\, du\,.
\end{aligned}$$

It follows from Proposition 4.1 that $\check{E}$ is stable for all α sufficiently close to $\pi/2$. Hence, if $\check{E}$ is unstable for some $\alpha > 0$, there exists a smallest $\alpha = \alpha_0$, $0 < \alpha_0 < \pi/2$, for which $\check{E} = E_0$ (correspondingly, θ_0, λ_0) is unstable. It then follows, by Proposition 4.2, that $\inf_{\eta\in V_0(\theta_0),\ \int\eta^2=1} Q(\theta_0,\eta) = 0$, hence there exists $\eta_0 \in V_0(\theta_0)$, $\eta_0 \ne 0$, such that $Q(\theta_0,\eta_0) = 0$. We will show that this is not the case.

By (3.4) and (3.5) we have

$$V_0(\theta_0) = \{\eta \in W_{1,2}[0,1] : \int_0^1 \eta\sin\theta_0 = 0\}$$

(6.2)

$$Q(\theta_0,\eta) = \int_0^1 (\eta'^2 - \lambda_0\sin\theta_0\eta^2) - (2\lambda_0/d)(\int_0^1 \cos\theta_0\eta)^2$$

$\inf_{\eta\in V(\theta_0),\ \int\eta^2=1} Q(\theta_0,\eta) = Q(\theta_0,\eta_0) = 0$ implies (see Proposition 3.2 and Equations (3.7a,b) that η_0 satisfies the following system for some $\rho_0 \in \mathbb{R}$:

$$\eta_0''(t) + \lambda_0\sin\theta_0(t)\eta_0(t) + \sigma_0\cos\theta_0(t) + \rho_0\sin\theta_0(t) = 0$$

(6.3) $$\eta_0(0) = \eta_0(1) = 0, \quad \int_0^1 \eta_0\sin\theta_0 = 0, \quad \eta_0 \neq 0$$

$$\sigma_0 = (2\lambda_0/d)\int_0^1 \eta_0\cos\theta_0 .$$

The equation $\eta'' + \lambda_0\sin\theta_0\eta = 0$ has the general solution

(6.4) $$\eta = c_0\theta_0' + c_1\theta_0'\gamma_0, \quad \gamma_0(t) = \int_0^t (1/\sin\theta_0(\tau))d\tau .$$

By using the method of variation of parameters one finds for the general solution of the differential equation in (6.3):

(6.5) $$\eta_0(t) = -(\sigma_0/2\lambda_0)t\theta_0'(t) - (\rho_0/\lambda_0) + c_0\theta_0'(t) + c_1\theta_0'(t)\gamma_0(t) .$$

$\eta_0(0) = \eta_0(1) = 0$ give, since $\theta_0'(0) = \theta_0'(1) := \kappa_0$

(6.6) $$c_0 = \rho_0/\lambda_0\kappa_0, \quad c_1 = \sigma_0/2\lambda_0\gamma_0(1) .$$

By the use of integration by parts one finds

(6.7) $$\int_0^1 \eta_0\cos\theta_0 = -(\sigma_0/2\lambda_0)(\sin\alpha_0 - d) + c_1(\gamma_0(1)\sin\alpha_0 - 1)$$

and since this must equal $(dc_0/2\lambda_0)$ by (6.3), one obtains

(6.8) $$-(\sigma_0/2\lambda_0)\sin\alpha_0 + c_1(\gamma_0(1)\sin\alpha_0 - 1) = 0 .$$

(6.8) together with (6.6) gives $\sigma_0 = 0$, $c_1 = 0$. Thus, we are left with

(6.9) $$\eta_0 = (\rho_0/\lambda_0\kappa_0)(\theta_0' - \kappa_0) .$$

The final condition $\int_0^1 \eta_0 \sin\theta_0 = 0$ yields

$$(6.10) \qquad (\rho_0/\lambda_0\kappa_0)(2\cos\alpha_0 - \kappa_0 d) = 0 .$$

Since $\rho_0 = 0$ implies $\eta_0 = 0$, we must have

$$0 = G(\alpha_0) := \cos\alpha_0 - \kappa_0 d/2 .$$

By (6.1) we have $\kappa_0 = \theta_0'(0) = (2\lambda_0 \sin\alpha_0)^{1/2}$, $d/2 = (2\lambda_0)^{-1/2} \int_{\alpha_0}^{\pi/2} \sin^{1/2}u\, du$,

hence $G(0) = 1$, $G(\pi/2) = 0$, $G'(\alpha) = -(1/2)\sin^{-1/2}\alpha\cos\alpha \int_{\alpha}^{\pi/2} \sin^{1/2}u < 0$ for $0 < \alpha < \pi/2$. Therefore no η_0, ρ_0 satisfying (6.3) exist, and the proof of Proposition 6.1 is complete.

We prove next:

<u>Proposition 6.2.</u> A 2-point extremal interpolant E (angle-constrained or free) with 2 or more inflection points is unstable.

<u>Proof</u>. If E has at least 2 inflection points then E contains the basic 2-point extremal E_1 (see Sec. 5). By Proposition 3.3 it suffices to prove that E_1^*, which is E_1 with angle-constraint, is unstable. We do this by exhibiting $\eta_1 \in V_0(\theta_1)$, $\eta_1 \neq 0$, for which $Q(\theta_1, \eta_1) = 0$. As in the preceding proof, this will be the case if for some $\rho_1 \in \mathbb{R}$:

$$\eta_1'' + \lambda_1\eta_1\sin\theta_1 + \sigma_1\cos\theta_1 + \rho_1\sin\theta_1 = 0$$

$$(6.12) \qquad \eta_1(0) = \eta_1(1) = 0, \quad \int_0^1 \eta_1\sin\theta_1 = 0, \quad \eta_1 \neq 0$$

$$\sigma_1 = (2\lambda_1/d) \int_0^1 \eta_1\cos\theta_1 .$$

This system is satisfied by

$$\rho_1 = 0, \quad \eta_1(t) = (1 - 2t)\theta_1'(t) .$$

Indeed, one computes

$$\eta_1'' + \lambda_1\eta_1\sin\theta_1' = -\sigma_1\cos\theta_1', \quad \sigma_1 = 4\lambda_1$$

$$\int_0^1 \eta_1\cos\theta_1 = 2d = d\sigma_1/2\lambda_1$$

$$\eta_1(0) = \eta_1(1) = \int_0^1 \eta_1\sin\theta_1 = 0 .$$

Here we have used $\theta_1(0) = \theta_1(1) = \theta_1'(0) = \theta_1'(1) = 0$.

7. Two-point angle-constrained interpolants with one inflection point.

If the 2-point angle-constrained extremal E contains one inflection point (either at one end or internally) then the problem of stability is more complex. If one proceeds from the inflection point 0 along E in one or the other direction to a terminal one traverses a proper subarc of the basic extremal E_1 (see Sec. 5). There is a point on E_1, close to the far terminal, - its precise location is given below - on which the stability of E depends. We call this point a stability focus. E may contain the right, the left or neither stability focus. We prove

Proposition 7.1. A 2-point angle-constrained extremal interpolant E with one inflection point is stable if E contains no stability focus.

Proof. E contains neither stability focus as one proceeds from the inflection point to one or the other terminal, hence is a subarc of another extremal $\hat{E}$, which is symmetric with respect to the inflection point and also contains no stability focus. By Proposition 3.3 it suffices to prove that the angle-constrained extremal $\hat{E}$ is stable. Let $t \mapsto \hat{\theta}(t) (0 \leq t \leq 1)$ be the normal representation of $\hat{E}$ and $(-b/2, -d/2)$, $(b/2, d/2)$, $(b > 0, d > 0)$ the coordinates of the terminals, with $\theta = 0$ along the positive x axis. We then have:

$$\frac{1}{2}\hat{\theta}'^2(t) = \hat{\lambda}\sin\hat{\theta}(t), \quad 0 \leq t \leq 1$$

$$\hat{\theta}(t) = \hat{\theta}(1 - t)$$

$$\hat{\theta}(0) = \alpha, \quad 0 < \alpha < \pi, \quad \hat{\theta}(1/2) = \hat{\theta}'(1/2) = 0$$

$$b = \int_0^1 \cos\hat{\theta} = 2(2\hat{\lambda})^{-1/2} \int_0^\alpha \cos u \cdot \sin^{-1/2} u \, du \tag{7.1}$$

$$d = \int_0^1 \sin\hat{\theta} = 2(2\hat{\lambda})^{-1/2} \int_0^\alpha \sin^{1/2} u \, du$$

$$(2\hat{\lambda})^{1/2} = 2 \int_0^\alpha \sin^{-1/2} u \, du \, .$$

It follows from Proposition 4.2 that $\hat{E}$ is stable for all α sufficiently small. Further if $\alpha = \pi$, E contains 2 inflection points, hence is unstable. Thus there is a smallest $\alpha = \alpha_*$, $0 < \alpha_* < \pi$, for which $\hat{E} = \hat{E}_*$ (correspondingly, θ_*, λ_*) is unstable. As one proceeds along this $\hat{E}_*$ from the inflection point to one of the terminals one reaches the (left or right) stability focus, mentioned in the statement of the proposition.

By Proposition 4.1, we are left to find α_* and θ_*, so that

$$\inf_{\eta \in V_0(\theta_*),\ \int \eta^2 = 1} Q(\theta_*, \eta) = 0,$$

where $\theta = \theta_*$ satisfies (7.1), with α replaced by α_*. By (3.4) and (3.5) we have

$$\begin{aligned} &V_0(\theta_*) = \{\eta \in W_{1,2}[0,1] : \int_0^1 \eta(b\cos\theta_* + d\sin\theta_*) = 0\} \\ (7.3)\qquad & \\ &Q(\theta_*,\eta) = \int_0^1 (\eta'^2 - \lambda_*\sin\theta_*\eta^2) - (2\lambda_*/d)\left(\int_0^1 \eta\cos\theta_*\right)^2 . \end{aligned}$$

The infimum 0 of $Q(\theta_*,\eta)$ is attained for $\eta = \eta_* \in V(\theta_*)$ if (see Equations (3.7a,b)) η_* satisfies the following system for some $\rho_* \in \mathbb{R}$:

$$\begin{aligned} &\eta_*'' + \lambda_*\eta_*\sin\theta_* + \sigma_*\cos\theta_* + \rho_*(b\cos\theta_* + d\sin\theta_*) = 0, \\ (7.4)\qquad &\eta_*(0) = \eta_*(1) = 0, \quad \int_0^1 \eta_*(b\cos\theta_* + d\sin\theta_*) = 0, \quad \eta_* \neq 0, \\ &\sigma_* = (2\lambda_*/d)\int_0^1 \eta_*\cos\theta_* . \end{aligned}$$

Using the general solution

$$\eta(t) = -\,[(\sigma_* + \rho_* b)/2\lambda_*]t\theta_*'(t) - \rho_* d/\lambda_* + c_0\theta_*'(t) + c_1\gamma_*(t)$$

$$(7.5)\qquad \gamma_*(t) = \begin{cases} -\,\theta_*'(t)\displaystyle\int_0^t (1/\sin\theta_*(\tau))d\tau & \text{for } 0 \le t < 1/2 \\ 2 & \text{for } t = 1/2 \\ \theta_*'(t)\displaystyle\int_t^1 (1/\sin\theta_*(\tau))d\tau & \text{for } 1/2 < t \le 1 \end{cases}$$

of the differential equation in (7.4), one finds after lengthy calculations,

$$(7.6)\qquad \eta_*(t) = (1 - 2t)\theta_*'(t) - \theta_*'(0), \quad \rho_* = \lambda_*\theta_*'(0)/d .$$

Using integration by parts and the relations, following from (7.1):

$$(7.7) \qquad 2\sin\alpha_* = \kappa_*^2/\lambda_*, \quad b = -2\kappa_*/\lambda_*, \quad \text{where} \quad \kappa_* = \theta_*'(0)$$

one obtains

$$(7.8) \qquad \sigma_* = (2\lambda_*/d) \int_0^1 \eta_* \cos\theta_* = 4\lambda_* - \lambda_*\kappa_* b/d \ .$$

Then one verifies readily that (7.6) solves the differential equation in (7.4); also

$\eta_*(0) = \eta_*(1) = 0$ and $\int_0^1 \eta_* \sin\theta_* = 2\cos\alpha_* - 2b - \kappa_* d$, hence

$$(7.9) \qquad \int_0^1 \eta_*(b \cos\theta_* + d \sin\theta_*) = -2\kappa_*^3/\lambda^2 + 2d \cos\alpha_* - \kappa_* d^2 \ .$$

Thus, all the conditions of (7.4) are satisfied if the quantity (7.9) is 0, or using (7.1) and (7.7) and the abbreviation

$$S(\alpha) = \int_0^\alpha \sin^{1/2} u \, du, \quad 0 \leq \alpha \leq \pi,$$

$$(7.10) \qquad F(\alpha_*) := \sin^{1/2}\alpha_* S^2(\alpha_*) + \cos\alpha_* S(\alpha_*) + 2\sin^{3/2}\alpha_* = 0 \ .$$

α_* is the unique root between $\pi/2$ and π of (7.10). Since $F(\pi/2) > 0$ and $F(\pi) < 0$, there is a root between $\pi/2$ and π, and since $F'(\alpha) < 0$, the root is unique (a rough estimate shows $\alpha \approx 171°$).

We have shown that $\hat{\theta}$, given by (7.1), with $\alpha < \alpha_*$ is stable and this completes the proof of Proposition 7.1.

The result in Proposition 7.1 is sharp because we have

Proposition 7.2. A 2-point angle-constrained extremal interpolant E with one inflection point is unstable if E contains the two stability foci.

Proof. In the proof of Proposition 7.1 it was seen that the extremal $\hat{E}_*$ whose terminals are the stability foci is unstable. By Proposition 3.3 E, which contains $\hat{E}_*$, is unstable.

There remains the case where the 2-point angle-constrained interpolant E contains one inflection point and one stability focus. We may assume that the normal representation θ of E is a solution of

$$\text{(7.11)}\qquad \begin{aligned} &\tfrac{1}{2}[\theta'(t)]^2 = \lambda\sin\theta(t), \quad 0 \le t \le 1 \\ &\theta(0) = \alpha, \quad \theta(1) = \beta \end{aligned}$$

for some $\lambda \in \mathbb{R}$, where

$$\text{(7.12)}\qquad \begin{aligned} &0 < \alpha < \alpha_* \le \beta < \pi \\ &\theta(t_0) = \theta'(t_0) = 0 \quad \text{for a unique} \quad t_0 . \end{aligned}$$

The numbers λ and t_0 are determined from the relations

$$\text{(7.13)}\qquad \sqrt{2\lambda} = \int_0^\alpha \sin^{-1/2}u\,du + \int_0^\beta \sin^{-1/2}u\,du, \quad \sqrt{2\lambda}t_0 = \int_0^\alpha \sin^{-1/2}u\,du .$$

It is seen that $t_0 \le 1/2$ and

$$\text{(7.14)}\qquad \theta(t_0 - \tau) = \theta(t_0 + \tau), \quad 0 \le \tau \le t_0 .$$

We now show that for each α, $0 < \alpha < \alpha_*$, there exists a unique $\beta = \beta_*(\alpha)$ such that the extremal E is stable if $\beta < \beta_*(\alpha)$ and is unstable if $\beta \ge \beta_*(\alpha)$. We say, the point on E for which θ has the value $\beta_*(\alpha)$ is <u>conjugate</u> to the point for which θ has the value α. As α approaches α_* (from below) $\beta_*(\alpha)$ approaches α_* from above, hence the stability foci of Proposition 7.1 are the special case of conjugate points where $\beta_*(\alpha) = \alpha$.

We now prove

<u>Proposition 7.3.</u> Suppose E is an angle-constrained extremal 2-point interpolant with normal representation $t \mapsto \theta(t)$, $0 \le t \le 1$, which contains one inflection point and for which $\theta(0) = \alpha$, $0 \le \alpha \le \alpha_*$ (see (7.10)), $\theta(1) = \beta \ge \alpha_*$. E is stable if and only if $\beta < \beta_*(\alpha)$, where $\beta_*(\alpha)$ is the unique root betw-en α_* and π of Equations (7.17) below. As α increases from 0 to α_*, $\beta_*(\alpha)$ strictly decreases from π to α_*.

__Proof.__ By Propositions 6.2 and 7.1 E is stable if $\beta < \alpha_*$ and unstable if $\beta = \pi$. Let $\beta_*(\alpha)$ denote the smallest value of β for which E is unstable and let θ_* be the normal representation of the extremal E_* for which $\theta_*(0) = \alpha$, $\theta_*(1) = \beta_*(\alpha)$. There must then exist $\eta_* \in V_0(\theta_*)$, $\int \eta_*^2 = 1$, such that

$$\inf_{\eta \in V_0(\theta_*), \int \eta^2 = 1} Q(\theta_*, \eta) = Q(\theta_*, \eta_*) = 0 \;. \tag{7.15}$$

As in the proof of Proposition 7.1, we have for η_* the system (7.4). The general solution of the differential equation in (7.4) is given by (7.5). One computes, using integration by parts,

$$\begin{gathered}\int_0^1 \gamma_* \cos\theta_* = 1, \quad \int_0^1 \gamma_* \sin\theta_* = 2/\theta_*'(1) - 2/\theta_*'(0) \\ \int_0^1 t\cos\theta_*(t)\cdot\theta_*'(t)dt = \sin\beta - d, \quad \int_0^1 t\sin\theta_*(t)\cdot\theta_*'(t)dt = b - \cos\beta \;.\end{gathered} \tag{7.16}$$

The four conditions $\eta_*(0) = \eta_*(1) = 0$, $\int_0^1 \eta_* \cos\theta_* = d\sigma_*/2\lambda$, $\int_0^1 (b\cos\theta_* + d\sin\theta_*)\eta_* = 0$ for $\eta_* \in V_0(\theta_*)$, and the condition $\eta_* \neq 0$, then lead to the equation

$$\begin{aligned} H(\alpha,\beta) := {} & (\sin\alpha\sin\beta)^{1/2}(S(\alpha) + S(\beta))^2 \\ & + (\sin^{1/2}\alpha\cos\beta + \sin^{1/2}\beta\cos\alpha)(S(\alpha) + S(\beta)) \\ & + 2(\sin\alpha\sin\beta)^{1/2}(\sin^{1/2}\alpha + \sin^{1/2}\beta)^2 = 0 \end{aligned} \tag{7.17}$$

for $\beta = \beta_*(\alpha)$. One finds that the function $\beta \mapsto H(\alpha,\beta)$ is strictly decreasing for $\alpha_* \leq \beta \leq \pi$. Also, if $0 < \alpha < \alpha_*$,

$$H(\alpha,\pi) < 0, \quad H(\alpha,\alpha) = 4F(\alpha)\sin^{-1/2}\alpha > 0, \tag{7.13}$$

where F in the function of (7.10) and $F(\alpha) > 0$ since $\alpha < \alpha_*$. It follows that $\beta_*(\alpha)$ is uniquely defined by (7.17). Then $\theta = \theta_*$, with $\theta_*(0) = \alpha$, $\theta_*(1) = \beta_*(\alpha)$, is the normal representation of an extremal E_*, for which there exists $\eta_* \in V_0(\theta_*)$, with $\eta_* = 0$, such that (7.15) holds. Therefore, E_* is unstable, and by Proposition 3.3, E is unstable if E contains E_*, i.e. if $\beta \geq \beta_*(\alpha)$.

The function $\alpha \mapsto \beta_*(\alpha)$ is nonincreasing. For if $\beta_*(\alpha_1) < \beta_*(\alpha_2)$ for $\alpha_2 > \alpha_1$, then the angle-constrained unstable extremal E_1 with $\theta_1(0) = \alpha_1$, $\theta_1(1) = \beta_*(\alpha_1)$, is contained in the extremal E_2 with $\theta_2(0) = \alpha_2$, $\theta_2(1) = \beta_*(\alpha_1)$, which is stable since $\beta_*(\alpha_1) < \beta_*(\alpha_2)$. This is a contradiction to Proposition 3.3. Actually, β_* is strictly decreasing, for if $\beta_*(\alpha_1) = \beta_*(\alpha_2)$ for $\alpha_2 > \alpha_1$ then $\beta_*(\alpha)$ is constant for $\alpha_1 \le \alpha \le \alpha_2$, which is impossible since the function β_* is analytic. Clearly, $\beta_*(0) = \pi$ and $\beta_*(\alpha_*) = \alpha_*$, thus the proposition is completely proved.

We proceed to give a geometric interpretation of conjugate points on a simple elastica curve. At the same time we obtain the precise range of angles that an arc of the elastica, which contains one inflection point, can make with the chord connecting the endpoints.[†]

Let E be the simple elastica of Proposition 7.3, $t \mapsto \theta(t)\,(0 \le t \le 1)$ its normal representation, $\theta(0) = \alpha$, $\theta(1) = \beta$, with $0 \le \alpha \le \beta \le \pi$, and let p_0, p_1 be the local vectors of the terminals of E. In the original interpolation problem the length of the vector $p_1 - p_0$ and the angles A,B that E makes with $p_1 - p_0$ at the endpoint are prescribed. More precisely, let A,B denote the angles in $(-\pi,\pi]$ from the vector $p_1 - p_0$ to the oriented curve E at p_0, p_1, respectively. We investigate the relationship between α,β and A,B. Clearly,

(7.19a) $$\alpha - \beta = A - B\ .$$

If the inflection point is taken as the origin of a cartesian coordinate system xy, with the positive x-axis along $\theta = 0$, then the point $(t,\theta(t))$ on E has coordinates

$$x = \int_{t_0}^{t} \cos\theta(\tau)d\tau = (2/\sqrt{2\lambda})\sin^{1/2}\theta(t)\,\mathrm{sgn}(t - t_0)$$

$$y = \int_{t_0}^{t} \sin\theta(\tau)d\tau = (1/\sqrt{2\lambda})S(\theta(t)\,\mathrm{sgn}(t - t_0)\ .$$

Expressing the slope of the vector $p_1 - p_0$, we obtain

(7.19b) $$[S(\alpha) + S(\beta)]/[2\sin^{1/2}\alpha + 2\sin^{1/2}\beta] = \tan(\alpha - A)\ .$$

[†] I wish to express my gratitude to Dr. D. D. Pence for his valuable assistance with this investigation. - M.G.

Since p_1 is above and to the right of p_0 it follows that

(7.19c) $$0 < \alpha - A \le \pi/2$$

For each pair (α,β), with $0 \le \alpha \le \beta \le \pi$, $\alpha + \beta > 0$, there is a unique pair (A,B) with $-\pi < A \le B < \pi$ determined by Equations (7.19a,b,c) (actually $A \ge -\pi/2$).

Let $B(A;\alpha,\beta)$ be the angle B for fixed A,α,β, and set

(7.20) $$B^*(A) = \sup_{0\le\alpha\le\beta\le\pi} B(A;\alpha,\beta) = B(A;\alpha_A,\beta_A) .$$

It is readily found that if $\alpha_A = 0$ then $A = -\pi/2$ and $\beta_A = \pi$, and if $\beta_A = \pi$ then $A = -\pi/2$ and $\alpha_A = 0$; also if $\alpha_A = \beta_A$ then $\alpha_A = \alpha_*$ (solution of (7.10)) and $A = B^*(A)$. We write

(7.21) $$A^* = B(A^*;\alpha_*,\alpha_*) = \sup_{0\le\alpha\le\beta\le\pi} B(A^*;\alpha,\beta)$$

($A^* \approx 99.5°$). It follows that if A is neither $-\pi/2$ nor A^* then the supremum in (7.20) is attained in the interior of the region $0 \le \alpha \le \beta \le \pi$. Thus, (α_A,β_A) make $B = A - \alpha + \beta$ a maximum under the side condition (7.19b). It follows that $\alpha = \alpha_A$, $\beta = \beta_A$ satisfy the equations

(7.22) $$\begin{aligned} 1 + \mu(\partial/\partial\alpha)[S(\alpha) + S(\beta) - 2\tan(\alpha - A)(\sin^{1/2}\alpha + \sin^{1/2}\beta)] &= 0 \\ -1 + \mu(\partial/\partial\beta)[S(\alpha) + S(\beta) - 2\tan(\alpha - A)(\sin^{1/2}\alpha + \sin^{1/2}\beta)] &= 0 . \end{aligned}$$

Elimination of the multiplier μ, and use of (7.19b) yield

(7.23) $$H(\alpha_A,\beta_A) = 0$$

where H is the function (7.13). Thus $\sup B(A;\alpha,\beta)$ is attained for conjugate values α_A,β_A. This is also true in the excluded cases $A = -\pi/2$, $A = A^*$ since $(0,\pi)$, (α_*,α_*) are conjugate pairs. Each conjugate pair (α,β) occurs in this characterization; for if α,β are used in (7.19b,c) a unique A is obtained for which $\alpha = \alpha_A$, $\beta = \beta_A$. We have proved

<u>Proposition 7.4.</u> Suppose A, $-\pi < A < \pi$, is such that there exists a simple elastica E with terminals p_0,p_1, which contains an inflection point and which makes the angle A with the vector $p_1 - p_0$ at p_0. Then the largest angle B that E can make with $p_1 - p_0$ at p_1 is obtained if p_0,p_1 are conjugate points of E. Conversely, each pair of conjugate points is characterized in this way.

We proceed to determine the range of angles A, B that a simple elastica with one inflection point can make with the chord joining the endpoints. Because of symmetry it suffices to determine the half where $A \leq B$, which we denote as $R_{A \leq B}$. If $0 \leq A \leq A^*$ (see (7.21)) then the interval $\{A,\ A \leq B \leq B^*(A)\}$ is in $R_{A \leq B}$ ($B^*(A)$ as in (7.20)). If $A > A^*$ then there is no $B \geq A$ such that $(A,B) \in R_{A \leq B}$; this follows from the above discussion. Let us assume now $A < 0$. By (7.19c), we have $A \geq -\pi/2$; so fix A, $0 < A \leq -\pi/2$. Substitute $B - A + \alpha$ for β in (7.19b), which then defines B as a function of α. It is easily found that $\partial B/\partial\alpha$ at $\alpha = 0$ is $+\infty$. B takes on its minimum $B_*(A)$ for $\alpha = 0$, hence by (7.19a,b)

$$S(B_*(A) - A) = 2 \sin^{1/2}(B_*(A) - A)\tan(-A) . \tag{7.24}$$

It is easy to see that the interval $\{A,\ B_*(A) \leq A \leq B^*(A)\}$ is in $R_{A \leq B}$. In summary, we have

$$R_{A \leq B} = \{-\pi/2 \leq A < 0;\ B_*(A) \leq B \leq B^*(A)\} \cup \{0 \leq A \leq A^*;\ A \leq B \leq B^*(A)\}.$$

Remark. The general $(A,B) \in R_{A \leq B} \cup R_{B \leq A}$ is the image of two pairs (α,β), hence arises from two distinct simple elastica $E_{(A,B)}$. If angle-constrained, no more than one of these is stable. There may be no stable elastica at all for $(A,B) \in R_{A \leq B} \cup R_{B \leq A}$. Thus, if $0 \leq A \leq A^*$, $B = B^*(A)$, then $B = B(A;\ \alpha_A, \beta_A)$ and there is a unique $E_{(A,B)}$, whose terminals are at the conjugate α_A, β_A. By Proposition 7.3, the angle-constrained $E_{(A,B)}$ is not stable. It seems probable that this happens only on the boundary of $R_{A \leq B} \cup R_{3 \leq A}$.

The last proposition of this section deals with 2-point interpolants with angle constraint at only one end.

Proposition 7.5. A 2-point extremal interpolant E which is angle-constrained at one terminal and free at the other is stable if and only if E contains no stability focus.

Proof. For the normal representation $t \mapsto \theta(t)$ $(0 \leq t \leq 1)$ of E we may assume

$$\frac{1}{2}\,\theta'^2(t) = \lambda \sin\theta(t), \quad 0 \le t \le 1$$

$$(7.25) \qquad \theta(0) = \theta'(0) = 0; \quad \theta(1) = \beta > 0 \ \text{prescribed}$$

$$\int_0^1 \cos\theta = b \ \text{ and } \int_0^1 \sin\theta = d \ \text{ prescribed .}$$

If β is sufficiently small then E is clearly stable. If E is stable for some $\beta_1 > 0$ then, by the Corollary to Proposition 3.4, E is stable for each $\beta < \beta_1$. On the other hand, E is not stable if $\beta = \pi$ since in this case E is unstable even if angle-constrained. It follows that there exists β_*, $0 < \beta_* < \pi$, such that E is stable for $\beta < \beta_*$, but unstable for $\beta > \beta_*$. By the same arguments as in the earlier part of this section we conclude that we have

$$(7.26) \qquad \inf_{\eta \in V_0(\theta), \int \eta^2 = 1} Q(\theta,\eta) = Q(\theta,\eta_*) = 0,$$

where

$$(7.27) \qquad V_0(\theta) = \{\eta \in W_{1,2}[0,1] : \eta(1) = 0, \int_0^1 \eta(b \cos\theta + d \sin\theta) = 0\} .$$

For η_* we have the conditions (compare (7.4)):

$$(7.28) \qquad \begin{gathered} \eta_*'' + \lambda\eta_* \sin\theta + \sigma\cos\theta + \rho(b \cos\theta + d \sin\theta) = 0 \\ \sigma = (2\lambda/d) \int_0^1 \eta_* \cos\theta, \quad \eta_*'(0) = 0, \quad \eta_*(1) = 0, \quad \eta_* \neq 0 . \end{gathered}$$

The condition $\eta_*'(0) = 0$ results from the fact that if $\eta = \eta_*$ minimizes $Q(\theta,\eta)$ then η_* must satisfy the free boundary condition $\eta_*'(0) = 0$. Proceeding as in the proof of Proposition 7.1, one finds

$$(7.29) \qquad \eta_*(t) = t\theta'(t) - \theta'(1)$$

is a solution provided β (which enters (7.28) through λ,

$(2\lambda)^{1/2} = \int_0^\beta \sin^{-1/2} u \, du)$ is a zero of F, cf. (7.10). Thus, $\beta_* = \alpha_*$, the previously found stability focus.

8. The stability function

In the two preceding sections the stability problem was settled for all extremal 2-point interpolants. Let E^* now be an extremal, interpolating a general $(n + 1)$-points configuration $\{p_0,p_1,\ldots,p_n\}$, and free at the terminals p_0,p_n. For ease of formulation we introduce the

Definition. A subarc E_i^* of E^* between two consecutive interior nodes p_{i-1},p_i $(2 \leq i \leq n - 1)$ is said to be proper if E_i^* contains no pair of conjugate points. The terminal arcs E_1^* and E_n^* are proper if they contain no stability focus.

By Propositions 3.4, 7.3 and 7.5, E^* is unstable if any of the subarcs E_i^* is not proper. We state this important result as

Proposition 8.1. A necessary condition for stability of an extremal interpolant with free terminals is that each arc between consecutive interpolation nodes be proper.

It should be observed that by assuming all arcs are proper we do not exclude the presence of inflection points. However we will exclude, with little loss of generality, *inflection points at the knots*. We say E^* is decomposable if p_m for some m between 1 and $n - 1$ is an inflection point, otherwise E^* is indecomposable. If E^* is decomposable then the subarcs E_a from p_0 to p_m and E_b from p_m to p_n are (free) extremal interpolants, and it is readily seen that E^* is stable or unstable if both E_a and E_b, are stable or unstable, respectively (the case where one of the E_a,E_b is stable, the other unstable, is omitted).

For indecomposable extremal interpolants E^* which satisfy the necessary condition of Proposition 8.1 we find a computable function U^* of $n - 1$ variables $(n+1$ is the number of interpolation nodes) with the property that E^* is stable if and only if U^* has a local minimum at the critical point corresponding to E^*.

Let $s \mapsto \theta^*(s)$ $(0 \leq s \leq s_n^*)$ be the normal representation of E^*, with interpolation nodes $0 = s_0^* < s_1^* < \ldots < s_n^*$. Then for uniquely defined $\lambda_1^*,\ldots,\lambda_n^*$, $\mu_1^*,\ldots,\mu_n^*$ we have

$$(8.1)\quad \begin{aligned} &\theta^{*\prime\prime}(s) + \lambda_i^* \sin\theta^*(s) - \mu_i^* \cos\theta^*(s) = 0, \\ &\frac{1}{2}\,\theta^{*\prime 2}(s) - \lambda_i^* \cos\theta^*(s) - \mu_i^* \sin\theta^*(s) = 0, \quad s_{i-1}^* \le s \le s_i^* \\ &\int_{s_{i-1}^*}^{s_i^*} \cos\theta^* = b_i, \quad \int_{s_{i-1}^*}^{s_i^*} \sin\theta^* = d_i, \qquad i = 1,\dots,n \\ &\theta^{*\prime}(0) = 0, \quad \theta^{*\prime}(s_n^*) = 0 . \end{aligned}$$

In addition to (8.1) we have the corner conditions

$$(8.2)\qquad \theta^{*\prime}_i(s^* - 0) = \theta^{*\prime}(s_i^* + 0), \quad i = 1,\dots,n-1$$

The potential energy for E^* is

$$(8.3)\qquad U_0(E^*) = \int_0^{s_n^*} \theta^{*\prime 2} = \sum_{i=1}^{n} 2(\lambda_i^* b_i + \mu_i^* d_i) .$$

We choose an arbitrary number $\delta > 0$, set $s_n^* + \delta = S$, and extend θ^* to the interval $[0,S]$ by setting $\theta^*(s) = \theta^*(s_n^*)$ for $s_n^* < s \le S$. Every function θ in this section is in the space $W_{1,2} = W_{1,2}[0,S]$ and is constant on some interval $[s_n,S]$, where $0 < s_n = s_n(\theta) < S$.

As stated above, we assume each subarc E_i^* $(i = 1,\dots,n)$ of E^* is proper and also that $\theta^{*\prime}(s_i) \ne 0$ for $i = 1,\dots,n-1$. We set $\theta^*(s_i^*) = \alpha_i^*$ $(i = 0,1,\dots,n)$. For every $(n-1)$-tuple $\alpha = (\alpha_1,\dots,\alpha_{n-1})$ sufficiently close to $\alpha^* = (\alpha_i^*,\dots,\alpha_{n-1}^*)$ the system

$$(8.4)\quad \begin{aligned} &\theta''(s) + \lambda_i \sin\theta(s) - \mu_i \cos\theta(s) = 0, \\ &\frac{1}{2}\,\theta'^2(s) - \lambda_i \cos\theta(s) - \mu_i \sin\theta(s) = 0, \quad s_{i-1} \le s \le s_i \\ &\int_{s_{i-1}}^{s_i} \cos\theta = b_i, \quad \int_{s_{i-1}}^{s_i} \sin\theta = d_i, \qquad i = 1,\dots,n \\ &\theta'(0) = 0, \quad \theta'(s_n) = 0, \\ &\theta(s_j) = \alpha_j, \qquad\qquad j = 1,\dots,n-1 \end{aligned}$$

has a unique solution $\theta \in W_{1,2}$, with $\lambda_i, \mu_i \in \mathbb{R}$, $0 = s_1 < s_2 < \dots < s_n < S$, in a sufficiently small preassigned neighborhood of θ^*. This follows readily from the fact that each of the arcs E_i^* is proper. (8.4) is system (8.1) with additional conditions $\theta(s_j) = \alpha_j$ replacing the conditions (8.2). We let θ denote the solution of (8.4), E_α the $\{p_0, p_1, \dots, p_n\}$ interpolant represented by θ_α. The potential energy for E_α is

$$(8.5) \qquad U_0(E_\alpha) = \int_0^{s_n} \theta_\alpha'^2 = \sum_{i=1}^{n} 2(\lambda_i b_i + \mu_i d_i) \; .$$

We now introduce the function

$$(8.6) \qquad U^*(\alpha) - U_0(E_\alpha)$$

and call it the <u>stability function</u> (associated with the extremal E^*). It is defined in a neighborhood of α^*. We prove

<u>Proposition 8.2.</u> There is a neighborhood $N(\alpha^*) \subset \mathbb{R}^{n-1}$ of α^* such that α^* is the unique critical point in $N(\alpha^*)$ of the function U^*.

<u>Proof.</u> Let $W^0_{1,2}$ denote the metric space of functions $\theta \in W^*_{1,2}$ which interpolate the points p_i at nodes $s_i = s_i(\theta)$, $\Delta = s_0 < s_1 < \dots < s_n < s^*$, with the metric

$$(8.7) \quad d^0(\theta_1, \theta_2) = \max_{i=1,\dots,n} |s_i(\theta_1) - s_i(\theta_2)| + |\theta_1(0) - \theta_2(0)| + \left\{\int_0^S (\theta_1' - \theta_2')^2\right\}^{1/2} .$$

We can choose $\delta^* > 0$ so that the following three conditions are satisfied: (i) θ^* is the only extremal in

$$(8.8i) \qquad N(\theta^*) = \{\theta \in W^0_{1,2} : d^0(\theta, \theta^*) \le \delta^*\};$$

(ii) for each α in

$$(8.8ii) \qquad N(\alpha^*) = \{\alpha \in \mathbb{R}^{n-1} : |\alpha - \alpha^*| \le \delta^*\}$$

system (8.4) has a unique solution $\theta_\alpha \in N(\theta^*)$ and each restriction $\theta_\alpha|[s_{i-1}, s_i]$ $(i = 1, \dots, n)$ is proper; (iii) for $j = 1, \dots, n-1$

$$(8.8iii) \qquad \operatorname{sgn}\theta_\alpha'(s_j - 0) = \operatorname{sgn}\theta_\alpha'(s_j + 0) = \operatorname{sgn}\theta^{*\prime}(s_j^*) \; .$$

To prove the proposition it suffices to show that $\alpha \in N(\alpha^*)$ is a critical point of $U^*(\alpha)$ if and only if $\alpha = \alpha^*$.

By (8.5) we have

$$U^*(\alpha) = \sum_{i=1}^{n} (\lambda_i b_i + \mu_i d_i) \tag{8.9}$$

where the $\lambda_i = \lambda_i(\alpha)$, $\mu_i = \mu_i(\alpha)$ are determined from the interpolation and end conditions:

$$\begin{aligned} B_i(\alpha_{i-1},\alpha_i,\lambda_i,\mu_i) &:= \int_{s_{i-1}}^{s_i} \cos\theta_\alpha - b_i = 0 \\ D_i(\alpha_{i-1},\alpha_i,\lambda_i,\mu_i) &:= \int_{s_{i-1}}^{s_i} \sin\theta_\alpha - d_i = 0, \quad i = 1,\ldots,n \\ E_1(\alpha_0,\lambda_1,\mu_1) &:= \lambda_1\cos\alpha_0 + \mu_1\sin\alpha_0 = 0 \\ E_n(\alpha_n,\lambda_n,\mu_n) &:= \lambda_n\cos\alpha_n + \mu_n\sin\alpha_n = 0 \,. \end{aligned} \tag{8.10}$$

We now seek critical points of U^* as a function of $\alpha = (\alpha_1,\ldots,\alpha_{n-1})$ and the accessory variables $\lambda_1,\ldots,\lambda_n$, $\mu_1,\ldots,\mu_n$, α_0, α_n, under the $2n + 2$ side conditions (8.10). If α is a critical point then there exist multipliers $\rho_i,\sigma_i (i = 1,\ldots,n)$ and ω_1,ω_n such that

$$\frac{\partial}{\partial\gamma}\left\{\sum_{k=1}^{n} (\lambda_k b_k + \mu_k d_k + \rho_k B_k + \sigma_k D_k) + \omega_1 E_1 + \omega_n E_n\right\} = 0$$

where γ stands for each of the variables α_i,λ_i,μ_i .

Let first i $(2 \le i \le n - 1)$ be such that E_i^* has no inflection point. Then $\mathrm{sgn}\theta^{*\prime}(s_{i-1}^*) = \mathrm{sgn}\theta^{*\prime}(s_i^*) = 1$, say, and, by (8.8iii), $\theta_\alpha'(s) > 0$ for $s_{i-1} \le s \le s_i$. Thus, using (8.4), we find

$$\begin{aligned} B_i(\alpha_{i-1},\alpha_i,\lambda_i,\mu_i) &= \int_{\alpha_{i-1}}^{\alpha_i} \kappa_i^{-1}(u)\cos u \, du - b_i \\ D_i(\alpha_{i-1},\alpha_i,\lambda_i,\mu_i) &= \int_{\alpha_{i-1}}^{\alpha_i} \kappa_i^{-1}(u)\sin u \, du - d_i, \end{aligned} \tag{8.12}$$

where

$$\kappa_i(u) = (2\lambda_i\cos u + 2\mu_i\sin u)^{1/2} \,. \tag{8.13}$$

Using $\gamma = \lambda_i$ and $\gamma = \mu_i$ in (8.11), one obtains

$$b_i - \rho_i \int_{\alpha_{i-1}}^{\alpha_i} \kappa_i^{-3}\cos^2 - \sigma_i \int_{\alpha_{i-1}}^{\alpha_i} \kappa_i^{-3}\sin\cdot\cos = 0$$

$$d_i - \rho_i \int_{\alpha_{i-1}}^{\alpha_i} \kappa_i^{-3}\cos\cdot\sin - \sigma_i \int_{\alpha_{i-1}}^{\alpha_i} \kappa_i^{-3}\sin^2 = 0$$

or, since $b_i = \int_{s_{i-1}}^{s_i} \cos\theta_\alpha = \int_{\alpha_{i-1}}^{\alpha_i} \kappa_i^{-1}\cos = \int_{\alpha_{i-1}}^{\alpha_i} \kappa_i^{-3}(2\lambda_i\cos + 2\mu_i\sin)\cos$:

(8.14)

$$(2\lambda_i - \rho_i) \int_{\alpha_{i-1}}^{\alpha_i} \kappa_i^{-3}\cos^2 + (2\mu_i - \sigma_i) \int_{\alpha_{i-1}}^{\alpha_i} \kappa_i^{-3}\cos\cdot\sin = 0$$

$$(2\lambda_i - \rho_i) \int_{\alpha_{i-1}}^{\alpha_i} \kappa_i^{-3}\cos\cdot\sin + (2\mu_i - \sigma_i) \int_{\alpha_{i-1}}^{\alpha_i} \kappa_i^{-3}\sin^2 = 0 .$$

By the Schwarz inequality

$$\left(\int_{\alpha_{i-1}}^{\alpha_i} \kappa_i^{-3}\cos\cdot\sin\right)^2 < \int_{\alpha_{i-1}}^{\alpha_i} \kappa_i^{-3}\cos^2 \int_{\alpha_{i-1}}^{\alpha_i} \kappa_i^{-3}\sin^2$$

(equality cannot hold), hence (8.14) gives

(8.15) $$\rho_i = 2\lambda_i, \quad \sigma_i = 2\mu_i .$$

If $i = 1$ then $\theta^{*\prime}(s_1^*) \neq 0$ (since E_1^* is proper), say $\theta^{*\prime}(s_1^*) > 0$, and also $\theta_\alpha'(s) > 0$ for $0 < s \leq s_1$, hence (8.12) holds for $i = 1$ (the integrals involved are improper). To avoid the divergent integrals in (8.14), we set

(8.16i) $$B_1 = \lambda_1 F_1 + \mu_1 G_1, \quad D_1 = \mu_1 F_1 - \lambda_1 G_1$$

where

$$F_1 = (\lambda_1^2 + \mu_1^2)^{-1}\left[\int_{\alpha_0}^{\alpha_1} \kappa_1^{-1}(\lambda_1\cos + \mu_1\sin) - \lambda_1 b_1 - \mu_1 d_1\right]$$

$$= \frac{1}{2}(\lambda_1^2 + \mu_1^2)^{-1}\left[\int_{\alpha_0}^{\alpha_1} \kappa_1 - \lambda_1 b_1 - \mu_1 d_1\right]$$

(8.16ii)

$$G_1 = (\lambda_1^2 + \mu_1^2)^{-1}\left[\int_{\alpha_0}^{\alpha_1} \kappa_1^{-1}(-\lambda_1\sin + \mu_1\cos) + \lambda_1 d_1 - \mu_1 b_1\right]$$

$$= (\lambda_1^2 + \mu_1^2)^{-1}[\kappa_1(\alpha_1) + \lambda_1 d_1 - \mu_1 b_1] .$$

Using these expressions in (8.11), one can carry out the differentiations with respect to $\gamma = \lambda_1$ and $\gamma = \mu_1$, and one obtains (8.15) for $i = 1$. The same result is obtained for $i = n$.

Finally if j $(2 \leq j \leq n - 2)$ is such that $\operatorname{sgn}\theta^{*\prime}(s_{j-1}^*) = -\operatorname{sgn}\theta^{*\prime}(s_j^*) = 1$, say, (hence E_j^* has an inflection point), then by (8.8iii), $\theta_\alpha'(s)$ also changes sign in (s_{j-1}, s_j), and (8.12) is replaced by

$$B_j(\alpha_{j-1},\alpha_j,\lambda_j,\mu_j) = \left(\int_{\alpha_{j-1}}^{\beta_j} - \int_{\beta_j}^{\alpha_j}\right)(\kappa_j^{-1}\cos) - b_j$$

(8.17)

$$D_j(\alpha_{j-1},\alpha_j,\lambda_j,\mu_j) = \left(\int_{\alpha_{j-1}}^{\beta_j} - \int_{\beta_j}^{\alpha_j}\right)(\kappa_j^{-1}\sin) - d_j$$

where $\kappa_j(\beta_j) = 0$, $\alpha_{j-1} < \beta_j$, $\beta_j > \alpha_j$. To differentiate the improper integrals one replaces the B_j, D_j by functions F_j, G_j analogous to (8.16), then (8.11) for $\gamma = \lambda_j$ and $\gamma = \mu_j$ again yields (8.15) for $i = j$. It should be observed that β_j depends on $\alpha_{j-1}, \alpha_j, \lambda_j, \mu_j$, but $\partial F_j/\partial\gamma$ and $\partial G_j/\partial\gamma$ do not contain terms $\partial\beta_j/\partial\gamma$. We have now established (8.15) for $i = 1,\ldots,n$.

We next choose α_i $(i = 1,\ldots,n - 1)$ for γ in (8.11) and obtain

$$(\rho_i\cos\alpha_i + \sigma_i\sin\alpha_i)\kappa_i^{-1}(\alpha_i) = (\rho_{i+1}\cos\alpha_i + \sigma_{i+1}\sin\alpha_i)\kappa_{i+1}^{-1}(\alpha_i)$$

or, using (8.13) and (8.15): $\kappa_i(\alpha_i) = \kappa_{i+1}(\alpha_i)$, i.e.

$$\theta'_\alpha(s_i - 0) = \theta'_\alpha(s_i + 0), \quad i = 1,\ldots,n-1 \; . \tag{8.18}$$

Furthermore, by (8.4), $\theta'_\alpha(0) = \theta'_\alpha(s_n) = 0$. Thus we have shown that if α is a critical point of $U^*(\alpha)$ then θ_α satisfies (8.1) and (8.2), hence $\theta_\alpha = \theta^*$, $\alpha = \alpha^*$. That conversely $U^*(\alpha^*) = U_0(E^*)$ is a critical value of U^* follows immediately from the fact that $U_0(E^*)$ is a stationary value of U_0. Proposition 8.2 is proved.

The stability function U^* attains a minimum in the compact set $N(\alpha^*)$, say

$$U^*(\alpha_{\min}) = \min_{\alpha \epsilon N(\alpha^*)} U^*(\alpha) \; .$$

If $\alpha_{\min}$ is a critical point of U^* (i.e. $\alpha_{\min}$ is in the interior of $N(\alpha^*)$) then, by the preceding proposition, $\alpha_{\min} = \alpha^*$ and E^* minimizes the potential energy U_0 among all E_α with $\alpha \epsilon N(\alpha^*)$. The theorem below will show that in this case E^* minimizes U_0 among all the $\{p_0,p_1,\ldots,p_n\}$- interpolants sufficiently close to E^*, hence that E^* is stable. On the other hand, if $U^*(\alpha^*)$ is not a local minimum of U^* then there are interpolants E_α arbitrarily close to E^* for which $U_0(E_\alpha) = U^*(\alpha) < U^*(\alpha^*) = U_0(E^*)$, hence E^* is unstable. Thus, we arrive at the following effective stability criterion:

<u>Theorem.</u> Suppose the indecomposable extremal interpolant $E^* = E_{\alpha^*}$ has only proper subarcs E^*_i. Then E^* is stable if and only if the stability function U^* has a local minimum at α^*.

<u>Proof.</u> The proof depends critically on the following result which we formulate as a lemma.

<u>Lemma.</u> There exists a neighborhood $N_0(\theta^*) \subset N(\theta^*)$ such that $U_0(C) \geq U_0(E_\alpha)$ for each C with normal representation $\theta \epsilon N_0(\theta^*)$. Here $\alpha = \{\alpha_1,\ldots,\alpha_{n-1}\}$, $\alpha_i = \theta(s_i(\theta))$.

<u>Proof of Lemma.</u> Since each internal (terminal) arc of E_α between consecutive interpolation nodes, if considered as a 2-point extremal interpolant with two (one) angle constraints, is stable it is true that $U_0(C) \geq U_0(E_\alpha)$ for C sufficiently

close to $\mathcal{E}_\alpha, \alpha$ fixed. The lemma asserts that this inequality holds in a neighborhood that is independent of α.

We may assume θ in the form (2.5):

$$\theta = \theta_\alpha + \varepsilon\eta + \varepsilon^2\xi \tag{8.19}$$

with $\eta \in V_0(\theta_\alpha)$, $\eta(s_i(\theta_\alpha)) = 0$ $(i = 1,\ldots,n-1)$, $d^0(0,\eta) \le 1$, $d^0(0,\xi) \le 1$. We also may assume $d_i = \int_{s_{i-1}(\theta_\alpha)}^{s_i(\theta_\alpha)} \sin\theta_\alpha \ne 0$ (the integral is independent of α), otherwise d_i should be replaced by b_i. Then by (2.10), (3.5)

$$\int_0^S \theta'^2 = \int_0^S \theta_\alpha'^2 + \varepsilon^2 Q(\theta_\alpha,\eta) + R(\varepsilon),$$

$$Q(\theta_\alpha,\eta) = \int_0^S (\eta'^2 - \frac{1}{2}\theta_\alpha'^2\eta^2) + 2\sum_{i=1}^n d_i^{-1} \int_{s_{i-1}(\theta_\alpha)}^{s_j(\theta_\alpha)} (\cos\theta_\alpha)\eta \int_{s_{i-1}(\theta_\alpha)}^{s_i(\theta_\alpha)} \theta_\alpha'\eta . \tag{8.20}$$

where $R(\varepsilon)/\varepsilon^2 \to 0$ as $\varepsilon \to 0$, uniformly for $\theta \in N(\theta^*)$, $\alpha \in N(\alpha^*)$. The mappings $\alpha \mapsto s_i(\theta_\alpha)$ $(i = 1,\ldots,n)$, $\alpha \mapsto \theta_\alpha$, from $N(\alpha^*)$ to $\mathbb{R}$, $N(\theta^*)$, respectively, are continuous, and so is the mapping

$$\alpha \mapsto \inf_{\eta\in V_0(\theta_\alpha),\, d^0(\theta,\eta)\le 1} Q(\theta_\alpha,\eta) := q_\alpha . \tag{8.21}$$

Since $q_\alpha > 0$ for each $\alpha \in N(\alpha^*)$ it follows that $q_* = \inf_{\alpha\in N(\alpha^*)} q_\alpha > 0$ and, by (8.19), $\int \theta'^2 \ge \int \theta_\alpha'^2 + \frac{1}{2}\varepsilon^2 Q(\theta_\alpha,\eta)$ for all sufficiently small ε, say $|\varepsilon| \le \varepsilon_0$. We can now choose the neighborhood $N_0(\theta^*) \subset N(\theta^*)$ so that $\theta \in N_0(\theta^*)$ may be represented in the form (8.19) with $|\varepsilon| \le \varepsilon_0$. Then $\int \theta'^2 \ge \int \theta_\alpha'^2$, which proves the lemma.

<u>Proof of the Theorem.</u> We need to prove only the sufficiency of the condition. Thus, we assume there exists $\varepsilon_1 > 0$ such that $U^*(\alpha^*) \le U^*(\alpha)$ for $|\alpha - \alpha^*| \le \varepsilon_1$. If the neighborhood $N_1(\theta^*) \subset N_0(\theta^*)$ is sufficiently small then $|\alpha - \alpha^*| \le \varepsilon_1$ for each $\theta \in N_1(\theta^*)$, $\alpha = \{\theta(s_i(\theta)\}$. Using the Lemma, we have

$$\int_0^S \theta^{*\prime 2} = U^*(\alpha^*) \le U^*(\alpha) = \int_0^S \theta_\alpha'^2 \le \int_0^S \theta'^2 ,$$

hence $U_0(E^*)$ is a local minimum.

We present two examples, which illustrate the effectiveness of the propositions in this section.

Example 1. Suppose we have the configuration $\{p_0, p_1, p_2\}$ where $p_0 = (0,0)$, $p_1 = (1,0)$, $p_2 = (1,d)$. Without loss we may assume $d \geq 1$. It is easy to see that, for each d, there is an extremal interpolant E^*, which makes the angle α^* with the vector $p_1 - p_0$ at p_1, where α^* varies from $\pi/4$ to 0 as d varies from 1 to ∞. Here the stability function U^* is a function of a single variable α, which has been computed by Dr. D. Pence. It is found that $U^*(\alpha^*)$ is a local minimum for each d. By the Theorem, the above E^* is a stable extremal.

Example 2. Suppose the configuration to be interpolated is $\{p_1, p_2, p_3, p_4\}$ with $p_1 = (a,0)$, $p_2 = (1,0)$, $p_3 = (0,1)$, $p_4 = (0,a)$, where $-\infty < a < 1$. This configuration with $a = 0.5$ was mentioned first in the note [5] as an example for which there is no interpolating elastica, and this claim was, without examination, repeated in many subsequent publications. However, there are interpolating elastica, for each a, in particular there is one which is symmetric with respect to the symmetry axis of the configuration. This can be seen as follows. Let C_β be the symmetric interpolant of $\{p_1, p_2, p_3, p_4\}$ which is uniquely defined by the following conditions. C_β has continuous slope; the arcs $C_{1\beta}, C_{2\beta}, C_{3\beta}$ between the interpolation nodes are simple elastica; $C_{1\beta}$ and $C_{3\beta}$ have curvature 0 at p_1 and p_4 respectively; the tangent vector along $C_{1\beta}$ turns through the angle β. Clearly, for $\beta = \pi$, the curvature at p_2 jumps from 0 to a negative value; and for some $\beta < \pi$ the curvature at p_2 jumps from a negative value to 0. Therefore there is some value (it is unique) β_*, $0 < \beta_* < \pi$, such that C_{β_*} has continuous curvature at p_2 (thus also at p_3, in fact everywhere), and this is an extremal interpolant of $\{p_1, p_2, p_3, p_4\}$. In this way, for each a, $-\infty < a < 1$, a unique extremal interpolant, E^*, is defined. We will see that each of these extremals in unstable. The results are based on computations carried out by Dr. D. Pence.

If $a \geq a^*$, where $a^* \approx -.27$, then the terminal arcs of E^* are improper, hence E^* is unstable by Proposition 8.1. If $a < a^*$ then the hypotheses of the above Theorem are satisfied. Instead of the stability function $U^*(\alpha) = U_0(E_{\alpha^*})$, $\alpha = (\alpha_1,\alpha_2)$, we use the function of one variable which is the restriction of $U^*(\alpha)$ to $\alpha_2 = 3\pi/4 + \alpha_1$ (i.e., we consider only symmetric perturbations E_α of E^*). The computed results show that α^* is not a minimum point of this function, hence E^* is not stable.

The question whether there are stable extremal interpolants for the configuration $\{p_1,\ldots,p_4\}$ (which would necessarily be nonsymmetric) remains open.

9. Stability of closed extremals interpolating regular polygons.

An extremal interpolant E with normal representation $s \mapsto \theta(s)$ $(0 \leq s \leq \bar{s})$ is said to be closed if

$$\theta(0) = \theta(\bar{s}), \quad \theta'(0) = \theta'(\bar{s}) \ . \tag{9.1}$$

Let $p_0, p_1, \ldots, p_{n-1}$ be the vertices of a regular n-gon $(n \geq 3)$. In [2, Sec. 8] it was shown that there exist closed extremals that interpolate the configuration $\{p_0, p_1, \ldots, p_{n-1}, p_n = p_0\}$. In particular, there is one, $\dot{E}_n$, which has no inflection points. Let $s \mapsto \dot{\theta}_n(s)$, $0 \leq s \leq n$, be its normal representation. Its total variation $Va(\dot{\theta}_n)$ is minimal, $Va(\dot{\theta}_n) = 2\pi$. We prove

Proposition 9.1. The closed extremal $\dot{E}_n$ $(n \geq 3)$ is stable.

Proof. The course $\dot{E}_n$ consists of n congruent arcs, each of length 1, and the increment of angle along each arc is $2\pi/n$. We write $\dot{\theta}$ for its normal representation and define $\dot{\theta}(s + n) = \dot{\theta}(s)$. Then

$$\dot{\theta}(s) = \dot{\theta}(s - 1) + 2\pi/n \ . \tag{9.2}$$

We assume $p_0 = (0,0)$, and set $p_k - p_{k-1} = (b_k, d_k)$ $(k = 1, \ldots, n)$ with $b_1 = 0$, $d_1 = d > 0$. Then

$$b_k = \int_{k-1}^{k} \cos\dot{\theta} = -d \sin(k - 1)2\pi/n, \quad d_k = \int_{k-1}^{k} \sin\dot{\theta} = d \cos(k - 1)2\pi/n \ . \tag{9.3}$$

Because of symmetry we have

$$\dot{\theta}(0) = \pi/2 - \pi/n, \quad \dot{\theta}(1/2) = \pi/2 \ . \tag{9.4}$$

Also,

$$\begin{gathered} \tfrac{1}{2}\dot{\theta}'^2(s) = \dot{\lambda}\sin\dot{\theta}(s), \quad 0 \leq s \leq n \\ (2\dot{\lambda})^{1/2} = 2\int_0^{\pi/n} \cos^{-1/2}u \, du, \quad (2\dot{\lambda})^{1/2}d = 2\int_0^{\pi/n} \cos^{1/2}u \, du \ . \end{gathered} \tag{9.5}$$

The quadratic form (3.5) becomes in this case

$$\begin{aligned} Q(\dot{\theta},\eta) = & \sum_{k=0}^{n-1} \int_0^1 [\eta'^2(t + k)dt - \tfrac{1}{2}\dot{\theta}'^2(t + k)\eta^2(t + k)]dt \\ & - 2 \sum_{k=0}^{n-1} (\dot{\lambda}/d_k)(\int_0^1 \eta(t + k)\cos\dot{\theta}(t + k)dt)^2 \ . \end{aligned}$$

By (9.2) and (9.5)

$$(1/d_k)\int_0^1 \eta(t+k)\cos\dot{\theta}(t+k)\,dt = (1/d)\int_0^1 \eta(t+k)\cos\dot{\theta}(t)\,dt,$$

thus

$$(9.6)\quad Q(\dot{\theta},\eta) = \sum_{k=0}^{n-1}\left\{\int_0^1 [\eta'^2(t+k)dt - \tfrac{1}{2}\dot{\theta}'^2(t)\eta^2(t+k)]dt - (2\dot{\lambda}/d)\left[\int_0^1 \eta(t+k)\cos\dot{\theta}(t)dt\right]^2\right\}.$$

This form is to be minimized on the space (3.4):

$$V_0(\dot{\theta}) = \{\eta \in \dot{W}_{1,2} : \int_{k-1}^{k} (b_k\cos\dot{\theta} + d_k\sin\dot{\theta}) = 0,\ k = 1,\dots,r\}.$$

Here $\dot{W}_{1,2}$ denotes the $W_{1,2}$ -space of functions of period n. Using (9.3), we find

$$(9.7)\quad V_0(\dot{\theta}) = \{\eta \in \dot{W}_{1,2} : \int_0^1 \eta(t+k)\sin\dot{\theta}(t)dt = 0,\ k = 0,1,\dots,n-1\}$$

Put $\eta(t+k) = \eta_k(t)\ (k = 0,1,\dots,n-1)$. Clearly $Q(\dot{\theta},\eta_k) = Q(\dot{\theta},\eta_0)$ and $\eta_k \in V_0(\dot{\theta})$ if $\eta_0 \in V_0(\dot{\theta})$. If $Q(\dot{\theta},\eta)$ attains its infimum for $\eta = \eta_0$, then also for $\eta = \tilde{\eta} = (1/n)(\eta_0 + \eta_1 + \dots + \eta_{n-1})$, and η has period 1. For $\tilde{\eta}$ of period 1 (9.5) becomes

$$(9.8)\quad (1/n)Q(\dot{\theta},\eta) = \int_0^1 (\eta'^2 - \dot{\lambda}\eta^2\sin\dot{\theta}) - (2n\dot{\lambda}/d)(\int_0^1 \eta\cos\dot{\theta})^2$$

and (9.7) requires $\int_0^1 \eta\sin\dot{\theta} = 0$. Thus, η must change sign in $(0,1)$ and we conclude

$$(9.9)\quad \int_0^1 \eta'^2/\eta^2 \geq \int_0^1 (d\sin 2\pi t/dt)^2 \Big/ \int_0^1 (\sin 2\pi t)^2 = 4\pi^2.$$

From (9.5) we have the estimates

$$(9.10)\quad \begin{gathered}(2\dot{\lambda})^{1/2} < (2\pi/n)\cos^{-1/2}\pi/n\\ d > \cos\pi/n.\end{gathered}$$

With (9.9), (9.10) substituted in (9.8), we find

$$(9.11)\quad (1/n)Q(\dot{\theta},\eta) \geq 4\pi^2[1 - (1/n)\tan^2\pi/n - 1/2n^2\cos\pi/n]\int_0^1 \eta^2,$$

thus $\eta \mapsto Q(\dot{\theta},\eta)$ is positive definite for $n \geq 3$. By Proposition 1, $\dot{\theta}$ is stable.

REFERENCES

[1] A.E.H. Love, A Treatise on the Mathematical Theory of Elasticity, 4th Ed. Cambridge Univ. Press, London, 1927.

[2] M. Golomb and J. Jerome, Nonlinear interpolating spline curves and equilibrium positions of thin elastic beams. To appear.

[3] M. A. Malcolm, On the computation of nonlinear spline functions, SIAM J. of Num. An. 14 (1977), 254-279.

[4] E. H. Lee and G. E. Forsythe, Variational study of nonlinear spline curves, SIAM Review 15 (1973), 120-133.

[5] G. Birkhoff, H. Burchard and D. Thomas, Nonlinear interpolation by splines, pseudosplines, and elastica. General Motors Research Laboratories Report 468, February 1965.

[6] J. W. Jerome, Smooth interpolating curves of prescribed length and minimum curvature, Proc. AMS 51 (1975), 62-66.

[7] S. D. Fisher and J. W. Jerome, Stable and unstable elastica equilibrium and the problem of minimum curvature. J. of Math. An. and Appl., 53 (1976), 367-376.

MG/ed

ON THE OPTIMIZATION OF THE MODIFIED MAXIMUM ENTROPY SPECTRUM OF LINEAR ADAPTIVE FILTERS

Jacob Benson and Leon Kotin

Systems Analysis Division
Plans, Programs & Analysis Directorate
US Army Communications Research & Development Command
Ft. Monmouth, New Jersey 07703

Abstract. Optimization of modified maximum entropy spectra of linear adaptive filters requires computing min $|H(z)|$, where

$$H(z) = \prod_{k=1}^{N} (z-z_k)(z-z_k^*), \; |z_k| < 1,$$

is the transfer function of a finite impulse response least mean square error whitening filter. A method for computing the relative minima is presented.

1. Introduction: the problem. Processors have been developed which may be used to estimate the instantaneous frequency of digital signals. One such processor takes the form of a linear adaptive filter [1] which represents a relatively simple adaptation algorithm. The major computational load lies in the calculation of the "modified maximum entropy spectrum" as defined in [1]. This spectrum is given by:

$$Q(\omega) = 1/\left|1 - \sum_{\ell=1}^{L} g_{\ell}{}^{*} \exp(-i\omega\ell)\right|^{2} \equiv {}^{1}/\left|H(z)\right|^{2}, \qquad z \equiv e^{-i\omega}.$$

Except for a scale factor, $H(z)$ is the transfer function of a finite impulse response LMSE (least mean square error) whitening filter. The purpose of this whitening filter is to remove the coloration of the input sequence. If the input sequence contains a very narrow band spectrum centered at ω_o, or equivalently if the input spectrum has a pole very close to the unit circle, then the whitening filter places a zero at the frequency, very close to the unit circle [2]. Essentially this means that it is of interest to find the relative minimum values of $|H(z)|$ on the unit circle.

In the formulation considered here, $H(z)$ is a real polynomial (i.e., has only real coefficients) of degree 2N with no real zeros.

The zeros then occur in complex conjugate pairs: $z_1, z_1^*, z_2, z_2^*, \ldots, z_N, z_N^*$. The problem may then be expressed geometrically as follows:

Given 2N points (the zeros of H(z)) located in the interior of the unit circle C symmetrically with respect to the x-axis, determine the relative minima of the product

$$(1) \qquad |H(z)| = |z-z_1||z-z_1^*|\ldots|z-z_N^*| = \prod_{k=1}^{N} |z-z_k||z-z_k^*|$$

of the distances from the 2N given points to the variable point z on C.

In this report, we express the product in terms of a real polynomial $g_N(x)$ of degree 2N whose coefficients are given in terms of the zeros $z_1,\ldots,z_N$. As a function of a real variable, in contrast to H(z), this polynomial can then be minimized by standard elementary techniques.

2. Analysis: the solution. Let us first represent the zeros z_k in rectangular coordinates:

$$(2) \qquad z_k = x_k + iy_k, \ k = 1,2,\ldots,N,$$

with

$$(3) \qquad r_k^2 \equiv |z_k|^2 = x_k^2 + y_k^2.$$

Since z lies on C,

(4) $$z = e^{i\theta} = \cos\theta + i\sin\theta$$

whence, from (1),

(5) $$|H(z)|^2 = \prod_{k=1}^{N} |(e^{i\theta}-z_k)(e^{i\theta}-z_k^*)|^2.$$

Then

(6) $$|H(z)|^2 = \prod_{k=1}^{N} |e^{2i\theta} - 2x_k e^{i\theta} + r_k^2|^2$$

$$= \Pi\ [(\cos 2\theta - 2x_k\cos\theta + r_k^2)^2 + (\sin 2\theta - 2x_k \sin\theta)^2]$$

$$= \Pi\ [\ 2r_k^2 \cos 2\theta - 4x_k(1 + r_k^2)\cos\theta + r_k^4 + 4x_k^2 + 1].$$

With the identity $\cos 2\theta \equiv 2\cos^2\theta - 1$ and the substitution

(7) $$x \equiv \cos\theta,$$

(6) represents a real polynomial of degree 2N in the real variable x:

(8) $$g_N(x) \equiv |H(z)|^2 = \prod_{k=1}^{N} [4r_k^2 x^2 - 4x_k(r_k^2+1)x + 4x_k^2 + (r_k^2-1)^2].$$

In more conventional summation form, this can be shown to be

(9) $$g_N(x) = \sum_{n=0}^{2N} \left(\sum_{\ell_1+\ldots+\ell_N=n} (a_{1\ell_1} a_{2\ell_2} \ldots a_{N\ell_N})\right) x^n$$

where $\ell_j = 0$, 1 or 2 and each $a_{j\ell_j}$ is given in terms of the corresponding zero $z_j = x_j + iy_j$ as follows:

(10) $$a_{j0} \equiv 4x_j^2 + (r_j^2 - 1)^2$$

$$a_{j1} \equiv -4x_j(r_j^2 + 1)$$

$$a_{j2} \equiv 4r_j^2,$$

with $j = 1,2,\ldots,N$.

The relative minima of $g_N(x)$, and hence of $|H(z)|$, can now be obtained by elementary methods from a knowledge of the zeros of

$$(11) \qquad dg_N(x)/dx = \sum_{n=1}^{2N} (\Sigma\, (a_{1\ell_1}\, a_{2\ell_2} \ldots a_{N\ell_N} \mid \ell_j = 0,1,2;\ \sum_j \ell_j = n))nx^{n-1}.$$

However, because of the substitution (7) and the subsequent fact that

$$(12) \qquad dg_N(x)/d\theta = -\sin\theta\, dg_N(x)/dx,$$

as many as two additional minimal points on C may have been excluded among these values. Thus the values $\theta = 0$ and π (i.e., $z = 1$ and -1) may give the additional minima $|H(\pm 1)|$.

3. Example. To illustrate the foregoing, consider the case $N = 2$. Then

$$(13) \qquad |H(z)| = |z-z_1||z-z_1^*||z-z_2||z-z_2^*|$$

and (9), with the help of (10) becomes

$$(14) \qquad g_2(x) = a_{10}\, a_{20} + (a_{10}\, a_{21} + a_{11}\, a_{20})x + (a_{10}\, a_{22} + a_{11}\, a_{21} + a_{12}a_{20})x^2$$

$$+ (a_{11}\, a_{22} + a_{12}\, a_{21})\, x^3 + a_{12}\, a_{22}\, x^4$$

$$= [4x_1^2 + (r_1^2 - 1)^2]\,[4x_2^2 + (r_2^2 - 1)^2]$$

$$-4[(4x_1^2 + (r_1^2-1)^2)x_2(r_2^2+1) + x_1(r_1^2+1)(4x_2^2 + (r_2^2-1)^2)]x$$

$$+4[(4x_1^2 + (r_1^2-1)^2)r_2^2 + 4x_1(r_1^2+1)x_2(r_2^2+1) + r_1^2(4x_2^2+(r_2^2-1)^2]x$$

$$-16[x_1(r_1^2 + 1)r_2^2 + r_1^2x_2\,(r_2^2 + 1)]x^3$$

$$+16\, r_1^2 r_2^2\, x^4$$

$$\equiv \sum_{n=0}^{4} b_n\, x^n.$$

Differentiating, we obtain

$$(15) \qquad dg_2/dx = \sum_{n=1}^{4} nb_n x^{n-1},$$

where the b_n's are defined in (14).

The case $N = 2$ now culminates in the following result. All of the relative minimum values of $g_2(x) = |H(z)|^2$ must occur at some of the (at most) three zeros of dg_2/dx in (15), except possibly for $g_2(\pm 1)$ which may also be minima.

Note that because of the linear factors x_1 and x_2 in the coefficients b_1 and b_3, if $x_1 \leqq 0$ and $x_2 \leqq 0$ it is clear that the nonreal extreme points are all in the left half z-plane; by symmetry, if $x_1 \geqq 0$ and $x_2 \geqq 0$ then the nonreal extreme points are all in the right half z-plane.

4. Summary. We now summarize the technique of obtaining minima of the 2N th-degree real polynomial $H(z)$ on the unit circle in the complex z-plane, given its 2N zeros $z_1, z_1^*, \ldots, z_N, z_N^*$ within the unit circle. We proceed as follows:

a. Using (9), find the real polynomial $g_N(x) \equiv |H(z)|^2$ in the real variable x. The coefficients of $g_N(x)$ are given in (9)-(10) in terms of the zeros $z_1, \ldots, z_N$.

b. Now find the zeros of $g_N' \equiv dg_N(x)/dx$, where this function is given in (11). Among these zeros will beall those which yield the relative minima of $g_N(x)$ in $-1 < x < 1$, and hence of $H(z)$ other than $z = \pm 1$. This can be done by means of elementary calculus, e.g., by determining the sign of g_N'' at each zero of g_N', etc. There are standard programs for doing all this on the computer.

c. Check the endpoints of the interval for minima. E.g., if $g_N'(1)$ or $y_N'(-1) > 0$, then the values $g_N(1)$ or $g_N(-1)$ will give relative minimum values of $|H(z)|^2$ at +1 or -1, respectively.

REFERENCES

[1] L.J. Griffiths, "Rapid Measurement of Digital Instantaneous Frequency", IEEE Trans. Acoustics, Speech, and Signal Processing, vol. ASSP-23, pp. 207-222, Apr 1975.

[2] R.J. Keeler and L.J. Griffiths, "Acoustic Doppler extraction by adaptive linear prediction filtering", unpublished.

TIME-OPTIMAL REJECTION SEQUENCING

Paul T. Boggs and Robert L. Launer
U. S. Army Research Office
Mathematics Division
Research Triangle Park, N. C. 27709

The U. S. Army selects soldiers for certain jobs by means of a computer program which matches an individual's characteristics and qualifications against a job "template". The program in question checks 18 qualification categories. The time for an individual to be checked within a single category is not fixed, but the average time for the population in each category is known. When an individual fails to qualify within a category, he is not checked in any other category. The frequency of failing to qualify (rejection rate) for each category is also known. The problem is to determine how to sequence the categories within the computer program to minimize the expected time to complete the qualification check. It is assumed that at least one of the rejection rates is positive.

Since the number of individuals to be screened is very large, a substantial amount of computer time can be saved by using the optimal ordering of the categories. An example is given at the end of this paper in which the running time is reduced by a factor of 22. Finally, it is noted that for 18 categories, there are $18! \approx 6.4 \times 10^{15}$ different orderings making enumeration totally impractical: At 10^6 comparisons per second, it would take over 200 years to determine the optimal ordering. In order to solve the problem, an expression is developed for the expected time to process an individual given an arbitrary arrangement of the categories. The optimal solution is then analytically derived. Our result is an extension of work described in [1] for slightly different systems.

Suppose that the screening device (program) consists of $n \geq 2$ screening categories. Let $T_i (i=1,2,\ldots,n)$ represent the average time to screen an individual in category i, given that the individual is not rejected prematurely from that category, and T_i' the average time given that the individual is rejected prematurely. Finally, let R_i be the rejection rate in category i, that is, the proportion of the (finite) population rejected in category i. It is desired to compute the expected time to screen an individual for an arbitrary arrangement of the categories.

Consider a simple system with two screening categories. An individual may be rejected in category 1 or category 2 or may successfully pass through the system without being rejected. The total average times are T_1', $T_1 + T_2'$ and $T_1 + T_2$ respectively. Then the average time to complete screening is the weighted sum of these average times; the weights corresponding to the probabilities of the three mutually exclusive events described above. That is:

$$E = T_1' R_1 + (T_1 + T_2')(1-R_1)(R_2) + (T_1 + T_2)(1-R_1)(1-R_2) \quad .$$

This easily extends to a system of n categories [2] which yields the expected time,

$$E = T_1' R_1 + (T_1 + T_2')(1-R_1)R_2 + \ldots . + (T_1 + \ldots + T_k') \prod_{i=1}^{k-1} (1-R_i)R_k + \ldots$$
$$+ (T_1 + \ldots + T_n') \prod_{i=1}^{n-1} (1-R_i)R_n + \left[\sum_{i=1}^{n} T_i \right] \left[\prod_{i=1}^{n} (1-R_i) \right] \quad .$$

Let E' represent the same quantity as E except that the kth and (k+1)th categories are permuted. The kth and (k+1)th terms are given, respectively, by the following:

$$(1) \quad (T_1 + T_2 + \ldots + T_{k-1} + T_{k+1}') \prod_{i=1}^{k-1} (1-R_i)R_{k+1} \quad , \text{ and}$$

$$(2) \quad (T_1 + T_2 + \dots + T_{k-1} + T_{k+1} + T_k') \prod_{i=1}^{k-1} (1-R_i)R_k\,(1-R_{k+1}) .$$

The change in the expected time, E, caused by this permutation of terms is E'-E; positive or negative values correspond, respectively, to an increase or decrease in E. It is easily seen that only the kth and (k+1)th terms of E and E' differ. The difference is, (from (1) and (2)),

$$(3) \quad E'-E = R_k\, R_{k+1} \prod_{i=1}^{k-1} (1-R_i) \left[\frac{T_{k+1}}{R_{k+1}} - \frac{T_k}{R_k} - (T_{k+1} - T_{k+1}') + (T_k - T_k')\right] .$$

The coefficient $R_k R_{k+1} \prod_{i=1}^{k-1} (1-R_i)$ does not depend on the order of the kth and (k+1)th terms of E and E'. Thus, the expected time E is decreased by permuting the terms if the bracket is negative. That is,

$$(4) \quad \frac{T_{k+1}}{R_{k+1}} - (T_{k+1} - T_{k+1}') < \frac{T_k}{R_k} - (T_k - T_k') .$$

Note that if the first 2 terms are permuted, then (3) becomes,

$$R_1\, R_2 \left[\frac{T_2}{R_2} - \frac{T_1}{R_1} - (T_2 - T_2') + (T_1 - T_1')\right] , \text{ and (4)}$$

becomes,

$$\frac{T_2}{R_2} - (T_2 - T_2') < \frac{T_1}{R_1} - (T_1 - T_1').$$

Thus the optimal ordering of the categories (to minimize E) is to compute the quantities

$$(5) \quad G_k = \frac{T_k}{R_k} - (T_k - T_k')$$

and order these beginning with the smallest and ending with the largest.

Ties in G_k are irrelevant or may be broken by considering other factors.

If $T_k - T_k' = 0$, then the G_k have an easily recognizable physical meaning. In that case, the optimality criterion is to order $G_k^{-1} = R_k/T_k$ beginning with the largest. The physical units (dimensions) associated with G_k^{-1} are rejection rates per time, and the criterion says to use the highest rejection rates per time in the ordering to minimize E.

As an example, a test problem was generated for the case when $T_k = T_k'$ consisting of T_i, R_i, i=1,...,18 where the T_i are uniformly distributed on (0,10) and R_i are uniformly distributed on (0,1). These are listed in Table 1. The optimal and least efficient expected times are 1.37 and 29.88 respectively which have a ratio of approximately 22. This example, although obviously not definitive, does indicate that a significant savings is possible.

REFERENCES

[1] R. Conway, W. Maxwell and R. Miller, Theory of Scheduling, Addison-Wesley, New York, 1967.

[2] P. Whittle, Probability, John Wiley, London, 1970.

Table 1

T_i	R_i
.1	.60
3.8	.84
1.8	.33
3.3	.19
7.0	.11
2.4	.32
6.1	.88
3.1	.31
.4	.23
4.8	.78
8.9	.70
7.1	.37
2.8	.19
8.1	.05
3.4	.95
9.1	.43
7.8	.81
5.6	.80

UNIVERSITY OF WISCONSIN - MADISON
MATHEMATICS RESEARCH CENTER

BAND MATRICES WITH TOEPLITZ INVERSES

T. N. E. Greville and W. F. Trench[†]

Technical Summary Report #

ABSTRACT

It is shown that a square band matrix $H = (h_{ij})$ with $h_{ij} = 0$ for $j - i > r$ and $i - j > s$, where $r + s$ is less than the order of the matrix, has a Toeplitz inverse if and only if it has a special structure characterized by two polynomials of degrees r and s, respectively.

AMS(MOS) Subject Classification: 15A09, 15A57

Key Words: Toeplitz matrix, Band matrix

Work Unit Number 2 - Other Mathematical Methods

[†] Department of Mathematics, Drexel University, Philadelphia, Pennsylvania 19104.

Sponsored by the United States Army under Contract No. DAAG29-75-C-0024.

BAND MATRICES WITH TOEPLITZ INVERSES

T. N. E. Greville and W. F. Trench[†]

1. Introduction. A Toeplitz matrix is a square matrix in which all the elements on any stripe are equal, where we follow Thrall and Tornheim [4] in defining a stripe as either the main diagonal or any diagonal line of elements parallel to it. More precisely, $T = (t_{ij})_{i,j=0}^{m}$ is Toeplitz if there is a sequence $\{\phi_\nu\}_{\nu=-m}^{m}$ such that $t_{ij} = \phi_{j-i}$ for $0 \leq i, j \leq m$. We shall call a square matrix $H = (h_{ij})_{i,j=0}^{m}$ a band matrix if there are nonnegative integers r and s less than the order of the matrix such that $h_{ij} = 0$ for $j - i > r$ and for $i - j > s$. We shall call such a matrix strictly banded if $r + s \leq m$. In this paper we show that a strictly banded matrix has a Toeplitz inverse if and only if it has a special structure characterized by two polynomials of degrees r and s , respectively.

Strictly banded matrices with Toeplitz inverses have been encountered by Trench [6] in the study of stationary time series and by Greville [2] in extending moving-weighted-average smoothing to the extremities of the data.

2. The main theorem. We shall prove the following:

Theorem 1. Let

$$H = (h_{ij})_{i,j=0}^{m}$$

be a matrix of order $m + 1$ over a field F , and suppose

(2.1) $$h_{ij} = 0 \quad \text{if} \quad j - i > r \quad \text{or} \quad i - j > s ,$$

where

(2.2) $$r \geq 0 , \quad s \geq 0 , \quad \text{and} \quad r + s \leq m .$$

[†] Department of Mathematics, Drexel University, Philadelphia, Pennsylvania 19104.

Sponsored by the United States Army under Contract No. DAAG29-75-C-0024.

Then H is the inverse of a Toeplitz matrix if and only if

$$(2.3)\qquad \sum_{j=0}^{m} h_{ij}x^j = \begin{cases} x^i\, A(x) \displaystyle\sum_{\mu=0}^{i} b_\mu x^{-\mu}, & 0 \le i \le s-1, \\ x^i\, A(x)\, B(1/x), & s \le i \le m-r, \\ x^i\, B(1/x) \displaystyle\sum_{\nu=0}^{m-i} a_\nu x^\nu, & m-r+1 \le i \le m, \end{cases}$$

where $a_0 b_0 \neq 0$,

$$(2.4)\qquad A(x) = \sum_{\nu=0}^{r} a_\nu x^\nu, \quad B(x) = \sum_{\mu=0}^{s} b_\mu x^\mu,$$

and $A(x)$ and $x^s B(1/x)$ are relatively prime.

3. Preliminary Observations and Results. A Toeplitz matrix is clearly persymmetric[1] (i.e., symmetric about its secondary diagonal), and it is well known that the inverse of a persymmetric matrix is persymmetric. Careful examination of H as defined by (2.3) reveals that it is also persymmetric; in fact, it is quasi-Toeplitz, in that h_{ij} is a function of $j-i$ alone except for those elements in the $s \times r$ submatrix in the upper left corner of H and the $r \times s$ submatrix in the lower right corner. That is, if we define $\theta_{-s}, \theta_{-s+1}, \ldots, \theta_r$ by

$$A(x)\, B(1/x) = \sum_{\nu=-s}^{r} \theta_\nu x^\nu,$$

then $h_{ij} = \theta_{j-i}$ except in these two corner submatrices.

The proof of the necessity part of Theorem 1 rests on the following lemma, which follows trivially from the last four equations of [5].

Lemma 1 (Trench). If $H = (h_{ij})_{i,j=0}^{m}$ is the inverse of a Toeplitz matrix and $h_{00} \neq 0$, then the elements h_{ij} $(1 \le i, j \le m)$ are determined in

[1] The term "persymmetric" is used in this sense by Wise [7], Trench [5], Huang and Cline [3], and others. Aitken [1] uses it to mean a Hankel matrix (i.e., $t_{ij} = \phi_{i+j}$).

terms of h_{i0} $(0 \leq i \leq m)$ and h_{0j} $(0 \leq j \leq m)$ by the recursion formula[2]

$$(3.1) \qquad h_{ij} = h_{i-1,j-1} + \frac{1}{h_{00}} (h_{i0}h_{0j} - h_{m-j+1,0}\, h_{0,m-i+1}), \qquad 1 \leq i,\ j \leq m.$$

It is also useful for the necessity proof to note that if H satisfies (2.3) and $H_i(x) = \sum_{j=0}^{m} h_{ij}x^j$, then, by inspection ,

$$\begin{aligned}
H_0(x) &= b_0\, A(x) , && \\
H_i(x) &= xH_{i-1}(x) + b_i\, A(x) , && 1 \leq i \leq s , \\
H_i(x) &= xH_{i-1}(x) , && s + 1 \leq i \leq m - r , \\
H_i(x) &= xH_{i-1}(x) - a_{m-i+1}x^{m+1}B(1/x), && m - r + 1 \leq i \leq m.
\end{aligned}$$

This means that

$$(3.2) \qquad h_{ij} = \begin{cases} h_{i-1,j-1} + a_j b_i , & 1 \leq i \leq s , \\ h_{i-1,j-1} , & s + 1 \leq i \leq m - r , \\ h_{i-1,j-1} - a_{m-i+1}\, b_{m-j+1} , & m - r + 1 \leq i \leq m , \end{cases}$$

where $1 \leq j \leq m$. Conversely, if

$$(3.3) \qquad h_{i0} = a_0 b_i \quad (0 \leq i \leq s), \qquad h_{0j} = b_0 a_j \quad (0 \leq j \leq r) ,$$

$$(3.4) \qquad h_{i0} = 0 \quad (i > s), \qquad h_{0j} = 0 \quad (j > r),$$

and h_{ij} $(1 \leq i,j \leq m)$ are computed from (3.2), then H will be of the form (2.3).

The proof of the sufficiency part of Theorem 1 rests on the following improved version of a result of Huang and Cline [3].

Lemma 2 (Huang and Cline). A nonsingular persymmetric matrix $H = (h_{ij})_{i,j=0}^{m}$ with $h_{00} \neq 0$, partitioned as

[2] Though this formula was known long before the publication of [3], it can also be derived from Lemma 2 below by invoking the persymmetry of both H and P as defined there.

(3.5) $$H = \begin{bmatrix} h_{00} & f^T \\ g & H_m \end{bmatrix}$$

has a Toeplitz inverse if and only if the matrix

(3.6) $$P = H_m - h_{00}^{-1} gf^T$$

is persymmetric.

Proof. Partition H^{-1} as

$$H^{-1} = \begin{bmatrix} t_{00} & u^T \\ v & T_m \end{bmatrix} ,$$

where t_{00} is a scalar. Since $HH^{-1} = I_{m+1}$, it is easy to verify that $PT_m = I_m$ under the hypotheses stated here. If H^{-1} is Toeplitz then so is T_m, and consequently $P = T_m^{-1}$ is persymmetric. Conversely, if P is persymmetric, then $T_m = P^{-1}$ is also. Since H^{-1} is persymmetric, Lemma 1 of Huang and Cline [3] implies that H^{-1} is Toeplitz.

In their statement of Lemma 2, Huang and Cline assumed that H_m is nonsingular. This is unnecessary.

4. Proof of Theorem 1. We begin the proof of Theorem 1 with the following lemma.

Lemma 3. Suppose $H = (h_{ij})_{i,j=0}^m$ is of the form (2.3), with $a_0 b_0 \neq 0$. Then H is nonsingular if and only if $A(x)$ and $x^s B(1/x)$ are relatively prime.

Proof. We assume without loss of generality that $a_r b_s \neq 0$. For sufficiency, we will show that if $A(x)$ and $x^s B(1/x)$ are relatively prime and

(4.1) $$\sum_{i=0}^m c_i H_i(x) \equiv 0 ,$$

then

(4.2) $$c_i = 0 , \quad 0 \leq i \leq m ;$$

this implies that the rows of H are linearly independent, and so H is nonsingular. From (2.3) and elementary manipulations, we can rewrite (4.1) as

$$(4.3)\qquad A(x)P(x) + A(x)x^{s}B(1/x)Q(x) + x^{m-r+1}B(1/x)R(x) \equiv 0 \ ,$$

where

$$(4.4)\qquad P(x) = \sum_{i=0}^{s-1} c_i \ \beta_i(x) \ ,$$

$$(4.5)\qquad Q(x) = \sum_{i=s}^{m-r} c_i \ x^{i-s} \ ,$$

and

$$(4.6)\qquad R(x) = \sum_{i=0}^{r-1} c_{i+m-r+1} \ \alpha_i(x) \ ,$$

with

$$(4.7)\qquad \beta_i(x) = \sum_{j=0}^{i} b_{i-j} \ x^j$$

and

$$(4.8)\qquad \alpha_i(x) = \sum_{j=i}^{r-1} a_{j-i} \ x^j \ .$$

Now suppose $A(x)$ and $x^sB(1/x)$ are relatively prime. Then, since $m - r + 1 > s$ by (2.2), and $A(x)$ and $x^sB(1/x)$ are not identically zero because $a_0b_0 \neq 0$, (4.3) implies that $A(x)$ divides $R(x)$ and $x^sB(1/x)$ divides $P(x)$. Therefore $R(x) \equiv 0$ and $P(x) \equiv 0$ because $\deg P(x) < \deg x^sB(1/x)$ and $\deg R(x) < \deg A(x)$.

Since $b_0 \neq 0$, it follows from (4.7) that the polynomials $\beta_i(x)$ for $0 \leq i \leq s - 1$ are linearly independent, and so (4.4) and $P(x) \equiv 0$ give $c_i = 0$ for $0 \leq i \leq s - 1$. Similarly, since $a_0 \neq 0$, the polynomials $\alpha_i(x)$ for $0 \leq i \leq r - 1$ are linearly independent by (4.8), and (4.6) and $R(x) \equiv 0$ give $c_i = 0$ for $m - r + 1 \leq i \leq m$.

Finally, replacing $P(x)$ and $R(x)$ by zero in (4.3) gives $Q(x) \equiv 0$, and so, by (4.5), $c_i = 0$ for $s \leq i \leq m - r$, and (4.2) is established.

The converse is equivalent to the assertion that H is singular if $A(x)$ and $x^sB(1/x)$ are not relatively prime. If $A(x)$ and $x^sB(1/x)$ have a nonconstant common factor, then they have a common zero ξ in some extension field $\tilde{F}$ of F. From (2.3),

$$\sum_{j=0}^{m} h_{ij}\xi^{j} = 0, \qquad 0 \le i \le m,$$

which implies that the columns of H are linearly dependent over $\tilde{F}$, and so H is singular as a matrix over $\tilde{F}$. Since nonsingularity of a matrix is invariant under field extension, H is singular over any field containing its coefficients, and so over F.

Proof of Theorem 1. For necessity, we assume that (2.1) and (2.2) hold and that $H = T^{-1}$, where $T = (\phi_{j-i})_{i,j=0}^{m}$. We first show that $h_{00} \neq 0$. Since $HT = TH = I_{m+1}$, we have

$$(4.9) \qquad \sum_{\nu=0}^{r} h_{0\nu}\,\phi_{j-\nu} = \delta_{0j}, \qquad 0 \le j \le m$$

and

$$(4.10) \qquad \sum_{\mu=0}^{s} h_{\mu 0}\,\phi_{j+\mu} = \delta_{0j}, \qquad -m \le j \le 0,$$

where δ_{0j} is a Kronecker symbol. Let p be the smallest integer such that $h_{0p} \neq 0$, and consider the quantity

$$(4.11) \qquad \Lambda = \sum_{\nu=0}^{r} h_{0\nu} \sum_{\mu=0}^{s} h_{\mu 0}\,\phi_{p+\mu-\nu}.$$

Since $h_{0\nu}$ vanishes for $\nu < p$ and (4.10) applies for $\nu \ge p$, (4.11) reduces to

$$\Lambda = h_{0p}.$$

On the other hand, reversing the order of summation in (4.11) gives

$$\Lambda = \sum_{\mu=0}^{s} h_{\mu 0} \sum_{\nu=0}^{r} h_{0\nu}\,\phi_{p+\mu-\nu},$$

which by (4.9) reduces to h_{0p} if $p = 0$, and vanishes if $p > 0$. Thus there is a contradiction unless $p = 0$, and consequently $h_{00} \neq 0$.

Now choose a_0 and b_0 so that $a_0 b_0 = h_{00}$, and define $a_1, \ldots, a_r$ and $b_1, \ldots, b_s$ to satisfy (3.3). By substituting (3.3) and (3.4) into (3.1), it is easy to verify that the latter reduces in this case to (3.2). Thus, the elements of H are determined by $a_0, a_1, \ldots, a_r$ and $b_0, b_1, \ldots, b_s$ in the same way as are the elements of a matrix of the form (2.3). Consequently, H is of the form (2.3), with $A(x)$ and $B(x)$ as in (2.4). Since H is nonsingular, $A(x)$ and $x^s B(1/x)$ are relatively prime, by Lemma 3. This proves necessity.

For sufficiency, let H be defined by (2.3) and (2.4) with $a_0 b_0 \neq 0$ and $A(x)$ and $x^s B(1/x)$ relatively prime, and let (2.1) and (2.2) hold. Then H is persymmetric, and, by Lemma 3, nonsingular. Let $P = (p_{ij})_{i,j=1}^{m}$ be the matrix in (3.6), and note that the numbering of the rows and columns starts with one rather than zero. In this case f and g in (3.5) are given by

$$f^T = (b_0 a_1, \ldots, b_0 a_r, 0, \ldots, 0) \quad \text{and} \quad g^T = (a_0 b_1, \ldots, a_0 b_s, 0, \ldots, 0) ,$$

so

$$p_{ij} = h_{ij} - b_i a_j = h_{i-1,j-1} , \quad 1 \le i \le s , \quad 1 \le j \le r$$

(see (3.2)), and

$$p_{ij} = h_{ij} \quad \text{if} \quad i > s \quad \text{or} \quad j > r .$$

The last two equations imply that P is the analog of H with the same polynomials $A(x)$ and $B(x)$, but with m decreased by one. Hence P is persymmetric. Therefore H^{-1} is Toeplitz, by Lemma 2.

5. Computation of H^{-1}. We close by showing how to find H^{-1} if H satisfies (2.3), where $a_0 b_0 \neq 0$ and $A(x)$ and $x^s B(1/x)$ are relatively prime, so that $H^{-1} = T = (\phi_{j-i})_{i,j=0}^{m}$ is a Toeplitz matrix. If $r = s = 0$, then H is diagonal and the inversion is trivial. If $s > 0$ and $r = 0$, then H and H^{-1} are lower triangular, so $\phi_j = 0$ if $j > 0$, and by looking at the first column of $TH = I_{m+1}$, we see that

$$\phi_0 = (a_0 b_0)^{-1}$$

and

$$\phi_{-j} = -b_0^{-1} \sum_{\mu=1}^{s} b_\mu \phi_{-j+\mu} , \qquad j \geq 1 .$$

A similar argument disposes of the case where $r > 0$ and $s = 0$. Now suppose $r \geq 1$, $s \geq 1$, and $a_r b_s \neq 0$. By looking at the first row of $HT = I_{m+1}$ and the first column of $TH = I_{m+1}$, we see that

$$(5.1) \qquad \sum_{\nu=0}^{r} a_\nu \phi_{j-\nu} = b_0^{-1} \delta_{j0} , \quad 0 \leq j \leq m ,$$

and

$$(5.2) \qquad \sum_{\mu=0}^{s} b_j \phi_{-j+\mu} = a_0^{-1} \delta_{j0} , \quad 0 \leq j \leq m .$$

In particular, (5.1) and (5.2) imply that the vector

$$\Phi = [\phi_{s-1}, \phi_{s-2}, \ldots, \phi_{-r}]^T$$

satisfies the system

$$(5.3) \qquad \begin{cases} \displaystyle\sum_{\nu=0}^{r} a_\nu \phi_{j-\nu} = b_0^{-1} \delta_{j0}, & 0 \leq j \leq s-1 , \\ \displaystyle\sum_{\mu=0}^{s} b_\mu \phi_{-j+\mu} = 0 , & 1 \leq j \leq r . \end{cases}$$

Therefore, if this system has only one solution, we can obtain Φ by solving it, and then compute the remaining elements of $\phi_m, \phi_{m-1}, \ldots, \phi_{-m}$ from (5.2) and (5.3); thus

$$\phi_j = -a_0^{-1} \sum_{\nu=1}^{r} a_\nu \phi_{j-\nu} , \qquad s \leq j \leq m ,$$

and

$$\phi_{-j} = -b_0^{-1} \sum_{\mu=1}^{s} b_\mu \phi_{-j+\mu}, \quad r < j \leq m .$$

If $K = (k_{ij})_{i,j=1}^{r+s}$ denotes the matrix of coefficients of the system (5.3), and

$$K_i(x) = \sum_{j=1}^{r+s} k_{ij} x^{j-1}$$

is the generating function of the elements of the ith row, then

$$(5.4) \qquad K_i(x) = \begin{cases} x^{i-1}\,A(x)\,, & 1 \le i \le s\,, \\ x^{i-1}\,B(1/x) & s < i \le r+s\,. \end{cases}$$

We shall show that K is nonsingular, which impl'es that (5.3) has a unique solution. If K were singular, then some nontrivial linear combination of its rows would equal the zero vector; thus, from (5.4) there would be constants $p_0, p_1, \ldots, p_{s-1}$ and $q_0, q_1, \ldots, q_{r-1}$, not all zero, such that

$$(5.5) \qquad A(x) \sum_{\nu=0}^{s-1} p_\nu\, x^\nu + x^s\, B(1/x) \sum_{\mu=0}^{r-1} q_\mu\, x^\mu \equiv 0\,.$$

But $A(x)$ and $x^s B(1/x)$ are relatively prime, so (5.5) implies that $A(x)$ divides $\sum_{\mu=0}^{r-1} q_\mu\, x^\mu$. Hence $q_0 = q_1 = \ldots = q_{r-1} = 0$, since $\deg A(x) = r$. This and (5.5) imply that $p_0 = p_1 = \ldots = p_{s-1} = 0$, a contradiction. Hence (5.3) has a unique solution.

A similar argument shows that, alternatively,

$$\Phi' = [\phi_s, \phi_{s-1}, \ldots, \phi_{-r+1}]^T$$

can be found by solving the system obtained by replacing the limits on j in (5.3) by $1 \le j \le s$ and $0 \le j \le r-1$, respectively.

It is now clear that the elements of H^{-1} do not depend on m, in that with $A(x)$ and $B(x)$ given, increasing m merely enlarges the sequence $\{\phi_\nu\}$ without changing the elements already determined. Thus, corresponding to every pair of polynomials $A(x)$ and $B(x)$ of degree r and s, respectively, with $a_0\, b_0 \neq 0$, such that $A(x)$ and $x^s B(1/x)$ are relatively prime, there is an infinite family of band matrices of the form (2.3) of all orders greater than or equal to $r + s$, all having Toeplitz inverses with elements taken from the sequence $\{\phi_\nu\}_{\nu=-\infty}^{\infty}$ that is the unique solution of (5.1) and (5.2).

REFERENCES

1. A. C. Aitken, Determinants and Matrices, 8th ed., Oliver and Boyd, Edinburgh, 1954.

2. T. N. E. Greville, Moving-weighted-average smoothing extended to the extremities of the data, MRC Technical Summary Report #1786, Mathematics Research Center, University of Wisconsin-Madison, August 1977.

3. N. M. Huang and R. E. Cline, Inversion of persymmetric matrices having Toeplitz inverses, J. Assoc. Comput. Mach., 19 (1972), 437-444.

4. R. M. Thrall and L. Tornheim, Vector Spaces and Matrices, Wiley, New York, 1957.

5. W. F. Trench, An algorithm for the inversion of finite Toeplitz matrices, J. Soc. Indust. Appl. Math., 12 (1964), 515-522.

6. __________, Weighting coefficients for the prediction of stationary time series from the finite past, SIAM J. Appl. Math., 15 (1967), 1502-1510.

7. J. Wise, The autocorrelation function and the spectral density function, Biometrika, 42 (1955), 151-159 .

STOCHASTIC THEORY OF ROTOR BLADE DYNAMICS

Y. K. Lin
Aeronautical and Astronautical Engineering Department
University of Illinois at Urbana-Champaign
Urbana, Illinois 61801

ABSTRACT. Turbulence in the atmosphere gives rise to stochastic terms in the governing equations for the blade motion. These terms appear in the coefficients for the unknown quantities, thus playing the role of parametric excitations, as well as in the inhomogeneous parts on the right hand sides of the equations. The parametric excitations affect the system stability, while the inhomogeneous excitations are important if the statistical properties of stable structural response are required in an analysis. Modeling turbulence as a random field, statistically stationary in time with very broad spectral densities, the structural response, treated as a vector, may be approximated by a Markov vector governed by the Itô type stochastic differential equations. The stochastic averaging scheme of Stratonovich is used to convert the original equations to the equivalent Itô equations, and the stability conditions are determined for various stochastic moments of the structural response, in terms of the turbulence spectral level, Lock number, and other structural and flight regime parameters.

I. INTRODUCTION. In the service life of a helicopter, numerous encounters with clear-air or thunderstorm turbulence can be expected. Furthermore, because of the very nature that lift is generated by blade rotation, some level of self-created turbulence is also unavoidable. Therefore, random turbulence in the atmosphere should be included in a realistic analysis.

Either natural or self-created turbulence may be modeled as a random process, statistically stationary in time. When viewed in a frame of reference which moves at the same velocity as the turbulence convection velocity, it may also be assumed as locally homogeneous. In this frame of reference, the turbulence appears to be a random pattern in space which changes very slowly in time. The well-known Taylor's hypothesis applies to the limiting case when the turbulence becomes a frozen pattern being transported at the convection velocity.

In normal operations, the speed of a rotor blade (rotation plus forward motion) is much greater than the convection speed of the turbulence. Therefore, the turbulence can become a rapidly changing random process with a very short correlation time (Ref. 1, p. 22), when it is observed on the moving blade (Ref. 2). When this correlation time is much shorther than the relaxation time of the blade system, the blade response to the turbulence excitations becomes very close to a Markov process (Ref. 1, p. 99), for which a large amount of information is available in the literature.

In this paper, a brief account will be given on the concept of Markov processes and the mathematical tools required to treat them. Next, the condition under which physical random phenomena may be approximated by Markov processes *will be discussed. Finally, the method will be applied to the* rotor blade problem. For a more detailed presentation, the reader is directed to Refs. 2 and 3.

II. MARKOV PROCESSES. A random process (generally vector valued) is a Markov process if its future probabilistic structure depends only on the present state and is independent of its past history. A sufficient condition for a random process X(t) to be Markovian is that the increments within arbitrary non-overlapping time intervals are statistically independent.

Among various Markov processes, the diffusive Markov process is governed by an Itô stochastic differential equation of the form (See, for example, Ref. 4),

$$dX_j = \varepsilon\, m_j(X,t) + \varepsilon^{1/2}\, \sigma_{jk}(X,t)\, dW_k(t) \tag{1}$$

where X_j are components of X(t), $W_k(t)$ are independent unit Wiener (Brownian motion) processes, and the usual summation convention of repeated indices in a product is implied. The coefficients m_j and σ_{jk} are called the drift and diffusion coefficients, respectively. The positive parameters ε and $\varepsilon^{1/2}$ on the right hand side of the equation are used to indicate the orders of magnitude for the two terms when both are equally significant. Every Wiener process has independent increments, thus is a Markov process itself.

The stochastic differential equation (1) is equivalent to the integral equation

$$X_j = X_j(t_o) + \varepsilon \int_{t_o}^{t} m_j(X,u)\, du + \varepsilon^{1/2} \int_{t_o}^{t} \sigma_{jk}(X,u)\, dW_k(u) \tag{2}$$

In the sense of Itô, the last integral in (2) is interpreted as a forward stochastic integral; i.e.,

$$\int_{t_o}^{t} \sigma_{jk}(X,u)\, dW_k(u) = \ell.\ i.\ m.\ \Sigma\, \sigma_{jk}(X,u)\, [W_k(u_{\ell+1}) - W_k(u_\ell)] \tag{3}$$

where ℓ. i. m. denotes a mean-square limit.

From Eq. (1), we can obtain another Itô equation for an arbitrary scalar function $\phi(X)$, provided that ϕ is twice differentiable with respect to the components of X:

$$d\phi = \left[\frac{\partial\phi}{\partial t} + \varepsilon\left(m_j \frac{\partial\phi}{\partial X_j} + 1/2\, \sigma_{j\ell}\, \sigma_{k\ell} \frac{\partial^2\phi}{\partial X_j\, \partial X_k}\right) \right] dt + \varepsilon^{1/2}\, \sigma_{jk} \frac{\partial\phi}{\partial X_j}\, dW_k \tag{4}$$

Equation (4) is, of course, reducible to (1) when $\phi = X_j$.

The relation (4) is known as Itô's differential rule. In particular, letting $\phi = X_r X_s$, we obtain

$$d(X_r X_s) = \varepsilon(m_r\, X_s + m_s\, X_r + 1/2\, \sigma_{r\ell}\, \sigma_{s\ell})\, dt + \varepsilon^{1/2}\, (\sigma_{rk}\, X_s + \sigma_{sk}\, X_r) dW_k \tag{5}$$

It is convenient to use the Itô stochastic integral and differential equation when dealing with diffusive Markov processes. The ensemble average of an Itô stochastic integral is zero, as can be deduced from the definition, Eq. (3). By the same token, the ensemble average of the last term in (1) or the last term in (4) must also be zero. This simplifies the calculation of the ensemble averages of X_j and $\phi(X)$ since they are governed by the conventional (and deterministic) differential equations

$$d\, E[X_j] \;/\; dt = \varepsilon\, E[m_j] \tag{6}$$

$$d\, E[\phi(X)] \;/\; dt = E[\frac{\partial\phi}{\partial t} + \varepsilon(m_j \frac{\partial\phi}{\partial X_j} + 1/2\, \sigma_{j\ell} \frac{\partial^2\phi}{\partial X_j\, \partial X_k})] \tag{7}$$

In some cases it may be possible to obtain the transition probability density $q(x,t \mid x_o,t_o)$ of $x(t)$, governed by the following parabolic differential equation, called the Fokker-Planck (or Kolmogorov forward) equation:

$$\frac{\partial q}{\partial t} + \varepsilon \frac{\partial}{\partial x_j} (m_j\, q) - \frac{\varepsilon}{2} \frac{\partial^2}{\partial x_j\, \partial x_k} (\sigma_{j\ell}\, \sigma_{k\ell} q) = 0 \tag{8}$$

subject to the initial condition

$$q(x,t_o | x_o,t_o) = \delta(x - x_o) \tag{9}$$

and some suitable boundary conditions. Note that the drift and diffusion coefficients in Eq. (1) appear in Eq. (8) but they are treated here as functions of the deterministic state variables x_j and time t. The transition probability is a conditional probability. If the drift and diffusion coefficients are independent of time, then q tends to the unconditional probability density p(x) of a stationary Markov process as the transition time $t - t_o$ increases. Being independent of time t, the stationary state probability density p is governed by

$$\frac{\partial}{\partial x_j} (m_j p) - (1/2) \frac{\partial^2}{\partial x_j\, \partial x_k} (\sigma_{j\ell}\, \sigma_{k\ell} p) = 0 \tag{10}$$

III. APPROXIMATION OF PHYSICAL PROCESSES BY MARKOV PROCESSES. Markov processes in general and the Wiener process in particular are mathematical idealizations. The real interest of an engineer lies in the approximation of real random phenomena by such processes. Obviously, for an approximation to be valid certain conditions must be satisfied. These will be discussed below.

Let the dynamical law of a physical problem be represented by

$$dX_j^*/dt = \varepsilon\, f_j(X^*, t) + \varepsilon^{1/2}\, g_{jk}\, (X^*, t)\, \xi_k(t) \tag{11}$$

where ξ_k are "physical" random processes. In the sequel, X^* will be referred to as the response, and ξ_k the excitations. It is useful to introduce the following definition of the correlation time of a random process, as a suitable measure of memory of the represented random phenomenon:

$$\tau_c = \int_0^\infty \tau |R(\tau)| \, dt \Bigg/ \int_0^\infty |R(\tau)| d\tau \tag{12}$$

where R is the correlation function of the random process, and τ is the time difference. It has been assumed tacitly that the process is at least weakly stationary and with a zero mean. If the relaxation time of the response is much longer than the correlation time of every excitation, then the response is expected to be close to a Markov process in some sense. Equation (11) suggests that the relaxation time of the response X^* is of the order ε^{-1}; therefore, X^* may be approximated by a Markov process if the correlation times of ξ_k are all much shorter than ε^{-1}.

We shall assume that a comparison between the relaxation time of the response and the correlation times of the excitations justifies the substitution of X^* by a Markov process X. The question now remains as to how Eq. (11) can be converted to an equivalent Itô equation (1). A fundamental difference between these two equations lies in the fact that $dW_k(t)$ are independent of X(t) whereas $\xi_k(t)$ are correlated with $X^*(t)$. To account for the correlation between the response and the excitations of a physical system at the same time instant t, Stratonovich has proposed a stochastic averaging procedure (Ref. 5, pp. 104-106) according to which

$$m_j = f_j(X,t) + \int_{-\infty}^{0} [\frac{\partial}{\partial X_\ell} g_{jk} (X,t)] \, g_{\ell r} (X, t + \tau) \\ E[\xi_k(t) \; \xi_r(t + \tau)]d\tau \tag{13}$$

$$\sigma_{j\ell} \, \sigma_{k\ell} = \int_{-\infty}^{\infty} g_{jr} (X,t) \, g_{ks} (X, t + \tau) \\ E[\xi_r(t) \; \xi_s (t + \tau)]d\tau \tag{14}$$

where distinction between X and X^* has been removed. Equation (14) gives the elements of the product matrix $\sigma\sigma'$ which are required in the computation of the ensemble average, Eq. (7), as well as the formulation of the Fokker-Planck equation, Eq. (8). Matrix σ itself is not needed in practice. Stratonovich further assumes that the parameter ε in Eq. (11) is small to justify taking time averages of Eqs. (13) and (14) for further simplification. The entire procedure, proposed initially on pure physical grounds, was later verified rigorously by Khasminskii in a limit theorem (Ref. 6).

In rotor blade dynamics, especially during high speed forward flights, an essential feature of the equations of motion is the periodic modulation of the coefficients which arises from the blade rotation. This unique feature would be lost if the coefficients were replaced by their time averages. Therefore, for high speed forward flights, Eqs. (13) and (14) must be retained without time averaging.

In the special case in which ξ_k are "physical" white noise processes; i.e.

$$E[\xi_k(t)\ \xi_r\ (t + \tau)] = 2\pi\ \Phi_{kr}\ \delta(\tau)$$

where Φ_{kr} are constants Eqs. (13) and (14) reduce to

$$m_j = f_j(X,t) + \pi\Phi_{kr} \frac{\partial}{\partial X_\ell} g_{jk}\ (X,t)\ g_{\ell r}(X,t) \tag{15}$$

$$\sigma_{j\ell}\ \sigma_{k\ell} = 2\pi\ \Phi_{rs}\ g_{jr}\ (X,t)\ g_{ks}\ (X,t) \tag{16}$$

These are the same results obtained independently by Wong and Zakai using a different approach (Ref. 7). The second term on the right hand side of Eq. (15) is called, some times, the Wong and Zakai correction.

IV. EQUATIONS OF MOTION. To illustrate how the theory of Markov process can be applied to the rotor blade dynamics, consider a linear model in which a blade undergoes the flapping and torsional motions. The following simplified assumptions proposed by Sissign and Kuczynski (Ref. 8) are adopted:

A. Structural Assumptions

(1) For the flapping motion, the blade is rigid, centrally hinged, and with elastic restraint at the hinge.

(2) For the torsional motion, the blade is elastic and the torsional angle varies spanwise linearly.

(3) The mass and elastic centers coincide along the 1/4 chord line.

B. Aerodynamic Assumptions

(1) Flow is incompressible and sectionally two-dimensional (i.e., the spanwise flow is negligible.)

(2) The aerodynamic forces can be computed from the steady state theory, except for the aerodynamic damping due to blade pitching for which the more accurate quasi-steady theory should be used.

(3) The lift slope is the same constant in the normal and reverse flows.

(4) Flow separation and stall do not occur.

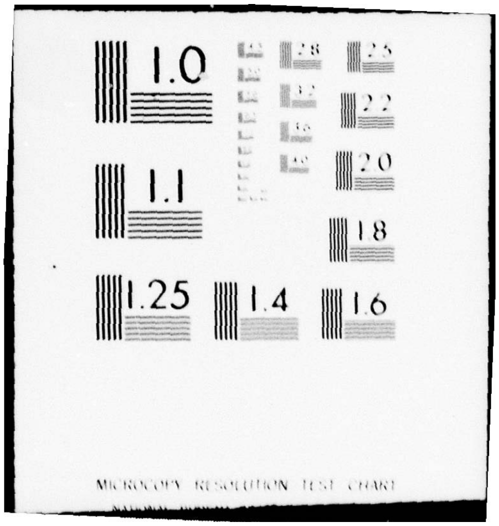
The equations of motion may be cast in a matrix form as follows (Refs. 2, 3)*:

*Only the homogeneous parts of the equations are given in Ref. 2 and 3. The inhomogeneous parts are added here for completeness.

AD-A064 433 ARMY RESEARCH OFFICE RESEARCH TRIANGLE PARK N C F/G 12/1
TRANSACTIONS OF THE CONFERENCE OF ARMY MATHEMATICIANS(24TH) HEL--ETC(U)
JAN 79

UNCLASSIFIED ARO-79-1 NL

5 OF 6
AD
A064433


1.0
2.8
2.5
3.2
2.2
3.6
4.0
2.0
1.1
1.8
1.25
1.4
1.6
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU

$$\begin{Bmatrix} \ddot{\beta} \\ \ddot{\alpha} \end{Bmatrix} + \frac{\gamma}{2} \begin{bmatrix} \overline{C} & 0 \\ 6Q\overline{\ell}_{r\dot{\beta}} & 6F\overline{C}_{\alpha} \end{bmatrix} \begin{Bmatrix} \dot{\beta} \\ \dot{\alpha} \end{Bmatrix} + \begin{bmatrix} p^2 + \frac{\gamma}{2}\overline{K} & -\frac{\gamma}{2}\overline{m}_{\alpha} \\ 3\gamma Q\overline{\ell}_{r\beta} & \omega_{\alpha}^{2} + 3\gamma Q\overline{K}_{\alpha} \end{bmatrix} \begin{Bmatrix} \beta \\ \alpha \end{Bmatrix} = \begin{Bmatrix} F_1 \\ F_2 \end{Bmatrix} \quad (17)$$

where α = torsion angle at the blade tip

β = flapping angle

γ = $R^4\rho ca/I_\beta$, blade Lock number

R = rotor radius

ρ = air density

c = blade chord

a = lift curve slope

I_β = flapping mass moment of inertia of a blade (kg-m^2)

Q = $cI_\beta/(4RI_\alpha)$

F = $(I_\beta/16I_\alpha)\ (c/R)^2$

I_α = feathering mass moment of inertia of a blade (kg-m^2)

$(\dot{\ })$ = derivative with respect to azimuth angle ψ; i.e. the non-dimensional time Ωt

Ω = blade angular velocity (rad/sec)

p = blade flapping frequency/blade angular velocity

and the coefficients $\overline{C}$, $\overline{K}$, ..., as well as the inhomogeneous excitations F_1 and F_2 are functions of the azimuth angle ψ, the forward flight velocity v, the inflow velocity w, and the turbulence velocity components. The tradictional procedure in helicopter dynamics is to non-dimensionalize velocities as fractions of the rotational speed of the blade tip $R\Omega$. Thus introducing

μ = $v/\Omega R$, advance ratio

λ = $w/\Omega R$, inflow ratio

η = horizontal turbulence velocity component along the flight path/ΩR

ξ = horizontal turbulence velocity component perpendicular to the flight path/ΩR

ν = vertical turbulence component/ΩR

the parameteric and non-parametric excitation terms in Eq. (17) are found to be

$$\begin{Bmatrix} \bar{C} \\ \bar{K} \\ \bar{m}_\alpha \\ \bar{C}_\alpha \\ \bar{K}_\alpha \\ \bar{\ell}_{r\beta} \\ \ell_{r\dot{\beta}} \end{Bmatrix} = \begin{bmatrix} C & C_\xi & C_\eta \\ K & K_\xi & K_\eta \\ m_\alpha & m_{\alpha\xi} & m_{\alpha\eta} \\ C_\alpha & C_{\alpha\xi} & C_{\alpha\eta} \\ K_\alpha & K_{\alpha\xi} & K_{\alpha\eta} \\ \ell_{r\beta} & \ell_{r\beta\xi} & \ell_{r\beta\eta} \\ \ell_{r\dot{\beta}} & \ell_{r\dot{\beta}\xi} & \ell_{r\dot{\beta}\eta} \end{bmatrix} \begin{Bmatrix} 1 \\ \xi \\ \eta \end{Bmatrix} \qquad (18)$$

$$\begin{Bmatrix} F_1 \\ F_2 \end{Bmatrix} = \begin{bmatrix} m_\lambda & m_\xi & m_\eta & m_\nu \\ n_\lambda & n_\xi & n_\eta & n_\nu \end{bmatrix} \begin{Bmatrix} \lambda \\ \xi \\ \eta \\ \nu \end{Bmatrix} \qquad (19)$$

The terms associated with the collective, linear, and cyclic pitch controls are not included in Eq. (19). The elements in the 7 x 3 matrix in (18) and those in the 2 x 4 matrix in (19) are functions of the azimuth angle ψ, the advance ratio μ, and the tip loss factor B. The dependence of these functions on ψ, although periodic as a whole, is different for different flow regions. The ones associated with the parametric excitations are tabulated in Ref. 2.

It is of interest to note that, for flapping and torsional motions, only the two horizontal components of the turbulence velocity contribute to the parametric excitations, whereas all three components appear in the inhomogeneous terms. Henceforth, our discussion will be restricted to system stability, and the inhomogeneous terms will be dropped. The mathematical technique to be used, however, is valid even when the inhomogeneous terms are included.

V. UNCOUPLED FLAPPING MOTION IN HOVERING FLIGHT. Letting α =0 in the first row of Eq. (17), we obtain the equation for uncoupled flapping motion:

$$\ddot{\beta} + \frac{\gamma}{2}\,\bar{C}\,\dot{\beta} + (p^2 + \frac{\gamma}{2}\,\bar{K})\beta = 0 \qquad (20)$$

In a hovering flight, $\mu = 0$, the mixed and reverse flow regions no longer exist. Then Eq. (20) is simplified to (Ref. 2)

$$\ddot{\beta} + [h + Z_1(t)]\,\dot{\beta} + [p^2 + Z_2(t)]\,\beta = 0 \tag{21}$$

where

$$Z_1(t) = (B^3\gamma/6)\,[\eta(t)\sin\psi + \xi(t)\cos\psi] \tag{22}$$

$$Z_2(t) = (B^3\gamma/6)\,[\eta(t)\cos\psi - \xi(t)\sin\psi \tag{23}$$

$$h = B^4\gamma/8$$

It is seen that even when $\xi(t)$ and $\eta(t)$ are stationary* random processes, $Z_1(t)$ and $Z_2(t)$ generally are non-stationary. However, if $\xi(t)$ and $\eta(t)$ are stationary, uncorrelated, and identically distributed, then it can be shown that $Z_1(t)$ and $Z_2(t)$ are also stationary. For such a case, the spectral densities and cross-spectral densities of $Z_1(t)$ and $Z_2(t)$ are

$$\overline{\Phi}_{11}(\omega) = \overline{\Phi}_{22}(\omega) = (B^6\gamma^2/72)\,[\overline{\Phi}_{\xi\xi}(\Omega - \omega) + \overline{\Phi}_{\xi\xi}(\Omega + \omega)] \tag{24}$$

$$\overline{\Phi}_{12}(\omega) = \overline{\Phi}_{21}(-\omega) = (i\,B^6\gamma^2/72)\,[\overline{\Phi}_{\xi\xi}(\Omega - \omega) - \overline{\Phi}_{\xi\xi}(\Omega + \omega)] \tag{25}$$

These are expressed in non-dimensional velocity squared per unit physical frequency (rad./sec). If furthermore, $\xi(t)$ and $\eta(t)$ are wide-band processes with a slowly varying spectral density in the frequency region of interest, then, in that region,

$$\overline{\Phi}_{11}(\omega) = \overline{\Phi}_{22}(\omega) \simeq (B^6\gamma^2/36)\,\overline{\Phi}_{\xi\xi}$$

and $Z_1(t)$ and $Z_2(t)$ also become nearly uncorrelated.

In the deterministic analysis where the effect of turbulence is not considered, the equation of motion in a hovering flight reduces to one with constant coefficients. Since damping exists in the system, the uncoupled flapping mode is always stable. Therefore, instability can only be caused by the turbulence.

The second order linear stochastic differential equation of the form, Eq. (21), has been considered by Ariaratnam and Tam (Ref. 9). In order to apply Stratonovich's stochastic averaging method, Eq. (21) must be transformed into two first order equations. Let

*Stationarity of a random process is interpreted in the weak sense here.

$$\beta = A(\psi) \cos \theta, \ \dot{\beta} = - A(\psi)\, p \sin \theta, \quad \theta = p\psi + \nu(\psi)$$

Eq. (21) may be replaced by two equations for $A(\psi)$ and $\nu(\psi)$:

$$\dot{A} = P(A, \theta, \psi) \sin \theta \tag{26}$$

$$\dot{\nu} = A^{-1} P(A, \theta, \psi) \cos \theta$$

where $P = [- (h + Z_1) \sin \theta + p^{-1} Z_2 \cos \theta] A$, and Z_1, Z_2 are treated as *functions* of non-dimensional time ψ. If the correlation times of Z_1 and Z_2 in the non-dimensional time scale are small compared with h^{-1}, then (A, ν) may be approximated by a vector Markov process. Carrying out both stochastic and time averaging, one obtains a pair of Itô equations:

$$dA = \varepsilon m_1 \, d\psi + \varepsilon^{1/2} [\sigma_{11} \, dW_1 + \sigma_{12} \, dW_2] \tag{27}$$

$$d\nu = \varepsilon m_2 \, d\psi + \varepsilon^{1/2} [\sigma_{21} \, dW_1 + \sigma_{22} \, dW_2]$$

where $W_1 (\psi)$ and $W_2(\psi)$ are independent Wiener processes, and

$$\varepsilon m_1 = - k_1 A = - \{(1/2)h - (\pi/8)[2\Phi_{11}(0) + 3\Phi_{11}(2p) + (3/p^2)\Phi_{22}(2p) + (6/p)\Psi_{21}(2p)]\} A$$

$$\varepsilon m_2 = - k_2 = - (\pi/4)[(2/p)\Phi_{21}(2p) - \Psi_{11}(2p) - (1/p^2)\Psi_{22}(2p)]$$

$$\varepsilon(\sigma\sigma')_{11} = k_3 A^2 = (\pi/4) [2\Phi_{11}(0) + \Phi_{11}(2p) + (1/p^2)\Phi_{22}(2p) + (2/p) \Phi_{21}(2p)] A^2$$

$$\varepsilon(\sigma\sigma')_{22} = k_4 = (\pi/4)[\Phi_{11}(2p) + (2/p^2) \Phi_{22}(0) + (1/p^2) \Phi_{22}(2p)]$$

$$\varepsilon(\sigma\sigma')_{12} = \varepsilon(\sigma\sigma')_{21} = - (\pi/2p) \Phi_{12}(0) A \tag{28}$$

The Φ_{jk} and Ψ_{jk} are cosine and sine spectral densities and cross-spectral densities of $Z_j(\psi)$ referred to the non-dimensional frequency ω/Ω, defined as

$$\Phi_{jk}(u) + i\, \Psi_{jk}(u) = \begin{cases} (1/\pi) \displaystyle\int_0^\infty E[Z_j(\psi)\, Z_k(\psi + \chi)] e^{-iu\chi} d\chi, & \text{if } j = k \\ (1/2\pi) \displaystyle\int_0^\infty E[Z_j(\psi)\, Z_k(\psi + \chi)] e^{-iu\chi} d\chi, & \text{if } j \neq k \end{cases} \tag{29}$$

The conversion from $\overline{\Phi}_{jk}$, Eqs. (24) and (25), to Φ_{jk} is given by

$$\overline{\Phi}_{jk}(\omega) = (1/\Omega) \; \Phi_{jk} \; (\omega/\Omega) \tag{30}$$

For the special case where $\xi(t)$ and $\eta(t)$ are stationary, uncorrelated, and identically distributed wide-band random processes, the transformation into $Z_1(t)$ and $Z_2(t)$ through Eqs. (22) and (23) preserves these properties. Then all the cross-spectral densities (either cosine or sine) are zero, and all the sine spectral densities are nearly zero. For such a case,

$$(\sigma\sigma')_{12} = (\sigma\sigma')_{21} = 0, \quad m_2 \simeq 0$$

and the two components, A and ν, of the vector Markov process become de-coupled. Furthermore, each component is itself a scalar Markov process. Of particular interest is the A-process since

$$A = (\beta^2 + p^{-2} \dot{\beta}^2)^{1/2}$$

therefore, boundedness in A implies boundedness in both β and $\dot{\beta}$.

The transition probability density $q_A(a, \psi | a_0, \psi_0)$ satisfies the Fokker-Planck euqtion (Appendix A):

$$\frac{\partial q_A}{\partial \psi} = k_1 \frac{\partial}{\partial a} (a q_A) + (k_3/2) \frac{\partial^2}{\partial a^2} (a^2 q_A) \tag{31}$$

where k_1 and k_3 are now

$$k_1 = (1/2)h - (\pi/8) \; [2\Phi_{11}(0) + 3\Phi_{11}(2p) + 3p^{-2} \; \Phi_{22}(2p)]$$

$$k_3 = (\pi/4) \; [2\Phi_{11}(0) + \Phi_{11}(2p) + p^{-2} \; \Phi_{22} \; (2p)]$$

The solution of Eq. (31) satisfying also the initial condition

$q_A(a, \psi_0 | a_0, \psi_0) = \delta(a - a_0)$ is

$$q_A = a^{-1} \; (2\pi k_3 \chi)^{-1/2} \; \exp\{-[\ell n(a/a_0) + k_3 k_5 \chi]^2/(2k_3\chi)\} \tag{32}$$

where $\chi = \psi - \psi_0$, $k_5 = (k_1/k_3) + 1/2$. From Eq. (32) the moments of the response amplitude A are evaluated to be

$$M_n(\chi) \equiv E[A^n(\psi)] = M_n(\psi_0) \; \exp \; \{-n[k_1 - (n-1) \; (k_3/2)] \; (\psi - \psi_0)\},$$

$$n = 1, 2, --- \tag{33}$$

from which the condition for stability in the nth moment is

$$2k_1 - (n-1)k_3 > 0 \tag{34}$$

Letting $h = B^4\gamma/8$ and

$$\Phi_{11} = \Phi_{22} = (B^6\ \gamma^2/36)\ \Phi_{\xi\xi}$$

the stability condition (34) becomes

$$\Phi_{\xi\xi} < 18\ \{\pi\ B^2\ \gamma[2(1 + p^{-2}) + n(3 + p^{-2})]\}^{-1} \tag{35}$$

The transition probability density, Eq. (32), represents the most complete solution attainable within the present framework of analysis. The stability condition, (34), for moments of arbitrary orders is also the most complete.

VI. HIGH FORWARD SPEED FLIGHTS. The results obtained for hovering flights are expected to hold also for small advance ratio μ. When μ is not small, the periodic variation in the coefficients $\overline{C}$, $\overline{K}$,.... can no longer be ignored. This suggests that, although the concept of stochastic average of Stratonovich is still sound for high forward speed flights, the time average portion of the whole procedure requires further consideration. One guideline for any stochastic analysis is that the results should be reducible to those of the corresponding deterministic analysis when the random terms are set to zero. The deterministic results available in the literature are those due to Sissingh (Ref. 10), and Peters and Hohenemser (Ref. 11) for the uncoupled flapping motion, and those due to Sissingh and Kuczynski (Ref. 8) for coupled flapping-torsional motion, all of which are based on un-averaged equations. In order to compare with these results the time average portion of the Stratonovich method cannot be used. Instead, use will be made of the Wong and Zakai corrections to obtain the drift and diffusion coefficients for the equivalent Itô equation (Ref. 7). Physically, this means that the random excitations $\xi(t)$ and $\eta(t)$ are replaced by "physical" white noise processes, and time averaging becomes unnecessary.

(1) Uncoupled Flapping Motion

Letting $X_1 = \beta$, and $X_2 = \dot{\beta}$, Eq. (20) may be replaced by two equations of the standard form (Eq. 11):

$$dX_j/d\psi = \varepsilon f_j(X,\ \psi) + \varepsilon^{1/2}\ g_{jk}(X,\ \psi)\ \xi_k(\psi)$$

where it can be identified that

$$\varepsilon f_1 = X_2$$

$$\varepsilon f_2 = -\ (p^2 + \frac{\gamma}{2} K)\ X_1 - \frac{\gamma}{2} C\ X_2$$

$$\varepsilon^{1/2} g_{11} = \varepsilon^{1/2} g_{12} = 0$$

$$\varepsilon^{1/2} g_{21} = -\ \frac{\gamma}{2}\ (K_\xi\ \ X_1 + C_\xi\ \ X_2)$$

$$\varepsilon^{1/2} g_{22} = - \frac{\gamma}{2} (K_\eta \ X_1 + C_\eta \ \ X_2) \tag{36}$$

$$\xi_1(\psi) = \xi(t)$$

$$\xi_2(\psi) \quad \eta(t)$$

By an application of Wong and Zakai corrections:

$$\varepsilon m = \overline{B} \ X \tag{37}$$

where

$$\overline{B} = \begin{bmatrix} 0 & 1 \\ & \\ -p^2 - \frac{\gamma}{2} K + C_{21} & - \frac{\gamma}{2} C + C_{22} \end{bmatrix} \tag{38}$$

$$C_{21} = \pi (\tfrac{\gamma}{2})^2 \ [C_\xi K_\xi \Phi_{\xi\xi} + (C_\xi K_\eta \Phi_{\xi\eta} + C_\eta K_\xi \Phi_{\eta\xi}) + C_\eta K_\eta \Phi_{\eta\eta}]$$

$$C_{22} = \pi (\tfrac{\gamma}{2})^2 \ [C_\xi{}^2 \Phi_{\xi\xi} + C_\eta C_\xi (\Phi_{\xi\eta} + \ \Phi_{\eta\xi}) + C_\eta{}^2 \Phi_{\eta\eta}]$$

The elements of the product diffusion matrix $\varepsilon(\sigma\sigma')$ are all zero, except

$$\varepsilon(\sigma\sigma')_{22} = (2\pi)\varepsilon[g_{21}^2 \ \Phi_{\xi\xi} + g_{21} \ g_{22}(\Phi_{\xi\eta} + \Phi_{\eta\xi}) + g_{22}^2 \ \ \Phi_{\eta\eta}]$$

$$= S_{31} \ X_1^2 + S_{32} \ X_1 \ X_2 + S_{33} \ X_2^2 \tag{40}$$

where

$$S_{31} = \pi \ \frac{\gamma^2}{2} \ [K_\xi^2 \ \ \Phi_{\xi\xi} + K_\eta K_\xi \ (\Phi_{\xi\eta} + \Phi_{\eta\xi}) + K_\eta{}^2 \ \Phi_{\eta\eta}]$$

$$S_{32} = \pi \ \frac{\gamma^2}{2} \ [2K_\xi \ \ C_\xi + (K_\xi \ C_\eta + K_\eta \ C_\xi)(\Phi_{\xi\eta} + \Phi_{\eta\xi}) + 2K_\eta C_\eta] \tag{41}$$

$$S_{33} = 2 \ C_{22}$$

It follows that the elements of the diffusion matrix $\varepsilon^{1/2} \ \sigma$ also are all zero, except

$$\varepsilon^{1/2} \ \sigma_{22} = [S_{31} \ X_1^2 + S_{32} \ X_1 \ X_2 + S_{33} \ X_2^2]^{1/2}$$

Thus the Itô equation for X is completely determined:

$$dX_j = \overline{B}_{jk} \ X_k \ d\psi + [S_{31} \ X_1^2 + S_{32} \ X_1 \ X_2 + S_{33} \ X_2^2]^{1/2} \ \delta_{2j} \ dW_j \tag{42}$$

where δ_{2j} is a Kronecker delta, and W_2 is a Wiener process.

The equations for the first moments of X_j are obtained by taking ensemble average of (42), resulting in

$$dE\,[X_j]/d\psi = \bar{B}_{jk}E[X_k] \tag{43}$$

Since the elements of matrix $\bar{B}$ are complicated periodic functions of ψ, a closed form solution is not obtainable. However, Eq. (43) is now deterministic, and it has the same form as the one solved by Peters and Hohenemser (Ref. 11) using a numerical procedure. Briefly, this technique involves numerical integration of Eq. (43) to obtain a transfer relationship between $E[X(0)]$ and $E[X(2\pi)]$ as follows:

$$E[X(2\pi)] = \bar{Q}E\,[X(0)] \tag{44}$$

where $\bar{Q}$ is the so-called Floquet transition matrix. The eigenvalue of $\bar{Q}$ having the largest absolute value determines the system stability, and the stability boundary is reached when this absolute value is equal to unity.

To determine the stability boundary for the second moments, we apply Itô's differential rule to obtain three stochastic equations for $Y_1 = X_1^2$, $Y_2 = X_1X_2$ and $Y_3 = X_2^2$, and then takes the ensemble averages. The results can be cast in a matrix form:

$$d\,E[Y]/d\psi = \bar{\bar{B}}\,E[Y] \tag{45}$$

where

$$[\bar{\bar{B}}] = \begin{bmatrix} 0 & 2 & 0 \\ \bar{B}_{21} & \bar{B}_{22} & 1 \\ S_{31} & 2\bar{B}_{21}+S_{32} & 2\bar{B}_{22}+S_{33} \end{bmatrix}$$

and $\bar{B}_{jk}$ are the (j, k) element of matrix $\bar{B}$.

Stability condition of Eq. (45) can again be investigated using the numerical method of the Floquet transition matrix.

Figure 1 shows the stability boundaries plotted on the γ vs p^2 plane for the first and second stochastic moments of the uncoupled flapping response, at an advance ratio $\mu = 2.4$ and corresponding to different turbulence levels. The two turbulence *velocity components*, parallel and perpendicular to the flight path, respectively, are assumed to be uncorrelated, but having the same spectra. The stability region is seen to be reduced by turbulence, the higher the turbulence level, the smaller the stable region. As expected, the 2nd moment stability region is always included in the 1st moment stability region. The same conclusion can be reached using the Schwarz inequality. Also included in the figure is the non-turbulence case as a baseline for comparison. This baseline agrees with the one previously obtained by Sissingh (Ref. 10) using an analog computer, and verified later by Peters and Hohenemser (Ref. 11) using the numerical Floquet matrix method. Since vector, $\{X_1, X_2\} = \{\beta, \dot{\beta}\}$ is treated as the response, the asymptotic stability for the first moment only assures that E[β] E [β] approach to zero. This condition is not as useful as the one obtained previously for the amplitude $A = (\beta^2 + p^{-2}\dot{\beta}^2)$ in the case of hovering flights. Under normal conditions, the first moment stability boundaries for the uncoupled flapping motion do not deviate much from the baseline. The one that is shown in Fig. 1 corresponds to an unusually high turbulence level of 0.00636 in order to demonstrate the general nature of such curves. The more useful stability boundaries are the ones for the second stochastic moments shown here for three spectral levels 0.00159, 0.00318, and 0.00636.

(2) Coupled Flapping-Torsional Motion

Mathematically, the additional degree of freedom does not change the basic nature of the problem, but algebraically, it is much more tedious. Therefore, only the final results will be illustrated herein.

Figure 2 shows the second moment stability boundaries for the coupled flapping-torsional motion, again assuming that the longitudinal and lateral turbulence velocities have the same spectral level but are uncorrelated. The response vector is now $\{X_1, X_2, X_3, X_4\} = \{\beta, \dot{\beta}, \alpha, \dot{\alpha}\}$ combining into ten second moments $E[X_1^2]$, $E[X_1X_2]$, $E[X_1X_3]$, $E[X_1X_4]$, $E[X_2^2]$, $E[X_2X_3]$, $E[X_2X_4]$, $E[X_3^2]$, $E[X_3X_4]$ and $E[X_4^2]$. The stability boundaries for the first moments are not shown in the figure for lack of practical importance. It is seen that the stability boundary of the coupled motion deviates substantially from the baseline of the non-turbulence case due to a rather low turbulence spectral level of $\phi_{\xi\xi} = 0.000318$ which can even occur from a natural geothermal source. All the stability boundaries shown in Fig. 2, including the baseline, are nearly straight on the $\gamma - p^2$ plane in the range of $p^2 > 1$, while rapid change takes place in the lower p^2 range. It is well known from the deterministic theory of parametric excitation involving the Mathieu-Hill type equations that a primary instability occurs at $p^2 = 0.25$. This accounts for the departure of stability boundaries in the region $0 < p^2 < 1$ from the general trend appearing in the region $p^2 > 1$. In helicopter dynamics, the flapping stiffness parameter p^2 of the blade is always greater than one. Therefore, results for $0 < p^2 < 1$ have no practical significance, but they are included here for completeness.

VII. CONCLUSIONS. The main conclusions reached in this exploratory study are summarized as follows:

(1) The flapping and torsional motions of a rotor blade operating in a three-dimensional turbulence field are governed by differential equations with periodic and random coefficients.

(2) For the uncoupled flapping motion or the coupled flapping torsional motion, the two horizontal turbulence components affect the system stability, but the vertical turbulence component contributes only to the non-parametric external force.

(3) The present analysis is valid for an arbitrary advance ratio, although a simplified assumption has been made that the random parametric excitations can be replaced by suitable white noise processes.

(4) Numerical calculations have confirmed the fact that the second moment stability guarantees the 1st moment stability.

(5) The 2nd moment stability boundaries of the uncoupled flapping motion deviate only slightly from the non-turbulence baseline, even under high level turbulence excitations. Thus, it remains very stable under normal circumstances.

(6) Contrary to the case of uncoupled flapping, the 2nd moment stability boundaries of the coupled flapping-torsional motion under realistic turbulence levels, differ significantly from the non-turbulence baseline.

ACKNOWLEDGEMENT. This paper is based on work supported by the U. S. Army Research Office under grant DAAG29-78-G-0039. Findings in the paper are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

REFERENCES

1. Stratonovich, R. L., Topics in the Theory of Random Noise, Vol. I, Translated by R. A. Silverman, Gordon and Breach, New York, N.Y., 1963.

2. Lin, Y.K., Fujimori, Y., and Ariaratnam, S.T., "Rotor Blade Stability in Turbulent Flows, Part I," to appear.

3. Fujimori, Y., Lin, Y.K., and Ariaratnam, S.T., "Rotor Blade Stability in Turbulent Flows, Part II," to appear.

4. Arnold, L., Stochastic Differential Equations: Theory and Applications, John Wiley and sons, Inc., New York, NY, 1974.

5. Stratonovich, R. L., Topics in the Theory of Random Noise, Vol, II, Translated by R. A. Silverman, Gordon and Breach, New York, NY, 1967.

6. Khasminskii, R. Z., "A Limit Theorem for the solution of Differential Equations with Random Right Hand Sides", Theory of Probability and Applications, Vol, 11, 1966, pp. 390-405.

7. Wong, E. and Zakai, M., "On the Relation Between Ordinary and Stochastic Equations", International Journal of Engineering Science, Vol. 3, No. 2, July, 1965, pp. 213-229.

8. Sissingh, G.J., and Kuczynski, W.A., "Investigations on the Effect of Blade Torsion of the Dynamics of the Flapping Motion", Journal of the American Helicopter Society, Vol. 15, No. 2, April, 1972, pp. 2-9.

9. Ariaratnam, S.T., and Tam, D. S. F., "Random Vibration and Stability of a Linear Parametrically Excited Oscillator", Zeitschrift fuer Angewandte Mathematik und Mechanik, (to appear)

10. Sissingh, G. J., "Dynamics of Rotors Operating at High Advance Ratios", Journal of the American Helicopter Society, Vol. 13, No. 3, July, 1968, pp. 56-63.

11. Peters, D.A., and Hohenemser, K.H., "Application of the Floquet Transition Matrix to the Problems of Lifting Rotor Stability", Journal of the American Helicopter Society, Vol. 16, No. 2, May, 1971, pp. 25-33.

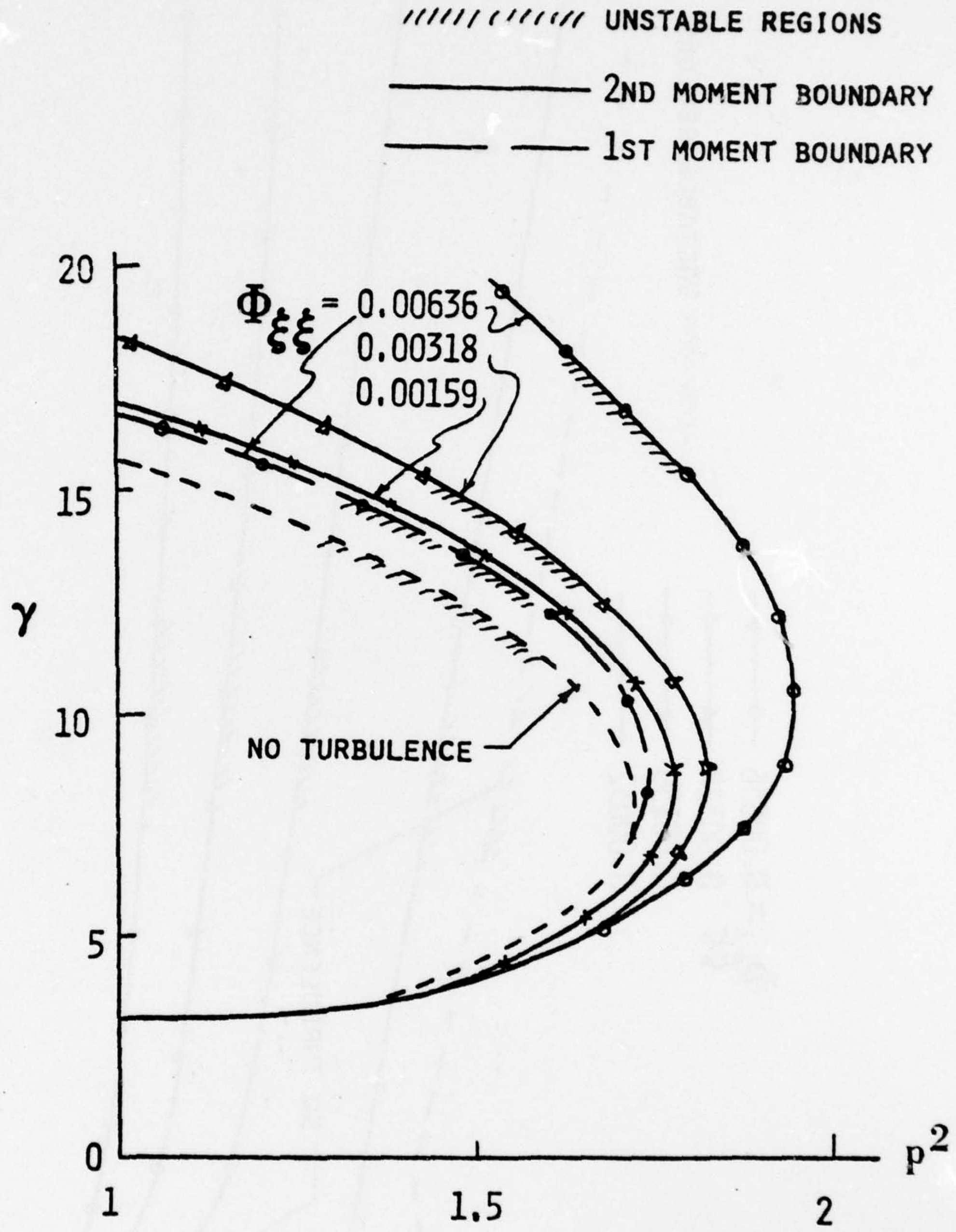


Fig. 1 First and Second Moment Stability Boundaries for Uncoupled Flapping Motion, $\mu = 2.4$, $\Phi_{\xi\xi} = \Phi_{\eta\eta}$, $\Phi_{\xi\eta} = 0$

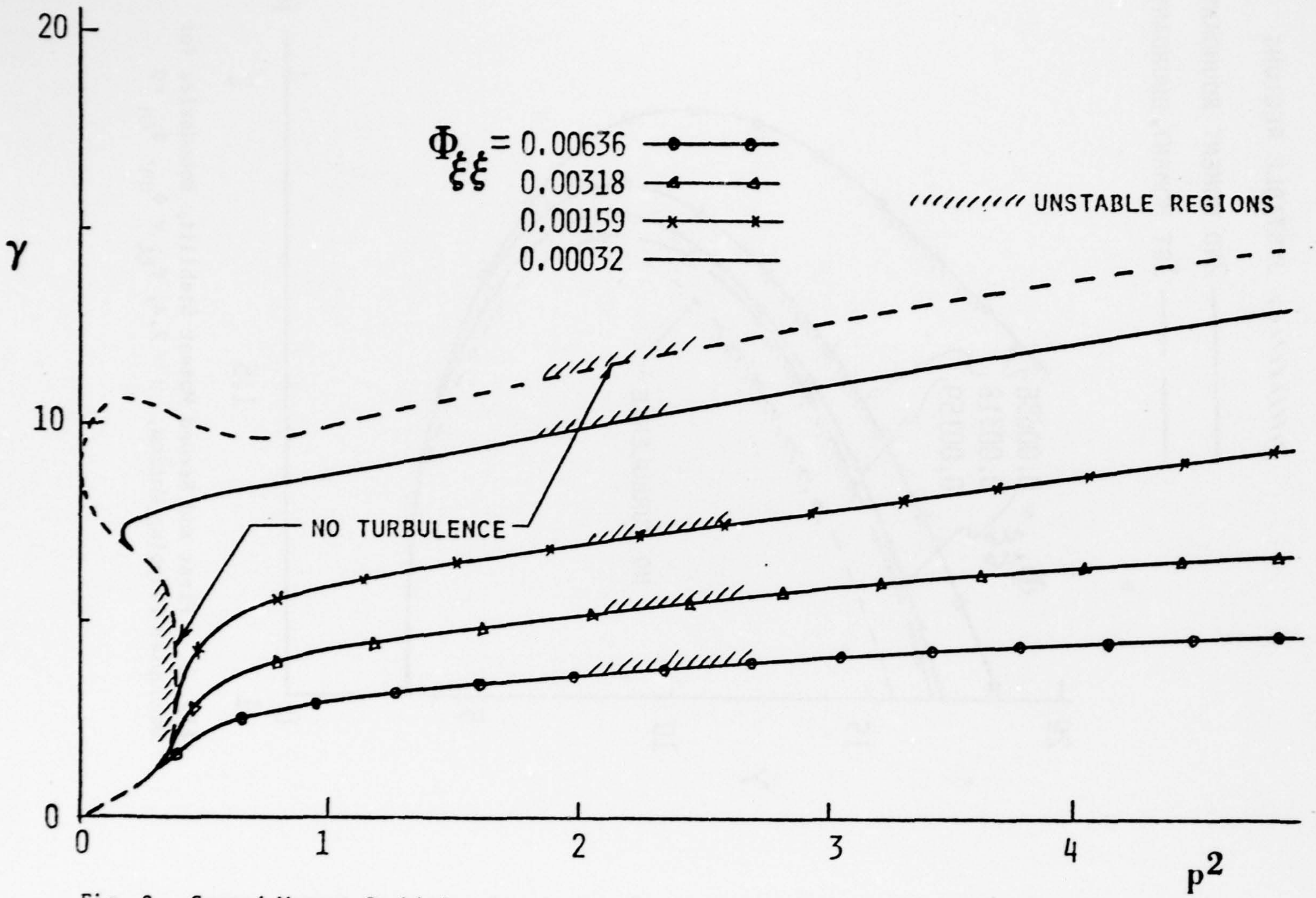


Fig. 2 Second Moment Stability Boundaries for Coupled Flapping-Torsional Motion, $\mu = 1.6$, $\omega_\alpha = 10$, $F = 0.24$, $Q = 15$, $\Phi_{\xi\xi} = \Phi_{\eta\eta}$, $\Phi_{\xi\eta} = 0$

MODELING AND ESTIMATION WITH BILINEAR STOCHASTIC SYSTEMS*

R.W. Brockett
Division of Applied Sciences
Harvard University
Cambridge, Massachusetts 02138

ABSTRACT. Over the past 10 to 15 years we have learned a great deal about the use of linear stochastic models in which the noise terms multiply the state variables. In this paper we survey some of this work. The discussion here is based on illustrative examples rather than theorems and proofs; for these the reader is given appropriate references.

I. INTRODUCTION. In practical engineering work the traditional Gauss-Markov process is certainly the workhorse. Occasional reference is made to other processes such as random telegraph waves, Poisson counters, etc. but even for such processes, it is very often second order statistics which are used. The relative tractability of the Gauss-Markov process is certainly an attractive feature, especially in a context where signal processing by linear systems is being contemplated, however it seems that these processes are relied on too much even if we take into account ones natural desire to get an answer in a finite amount of time.

The main point of this paper is that there is a more general class of processes which are, for some purposes, as easy or easier to deal with and which model some phenomena with more accuracy. The processes we will be concerned with are generated from two types of standard processes

(a) W = Wiener process of variance 1

(b) N = Poisson counter of rate λ

The properties of the Wiener process (the integral of white noise) are well known and explained very well in books such as Wong [9]. The Poisson counter of rate λ is an integer valued process which has the property that the expected value of $N(t)-\lambda t$ is zero and the probability that N(t) is r+k given that N(s) = r is

$$p[N(t) = r+k|N(s) = r] = \frac{1}{k!} \lambda^k (t-s)^k e^{-\lambda(t-s)}$$

These building blocks in conjunction with differential equations give rise to a rich and varied class of stochastic processes. In addition to the Gauss-Markov models given by the Itô equation

$$dx = Axdt + \sum_{i=1}^{r} b_i dw_i; \qquad y = cx$$

and the general finite-state, continuous-time jump process which we can model as

$$dx = \sum_{i=1}^{r} A_i x dN_i$$

*Partially supported by the Army Research Office under Grant DAAG29-76-C-0139, the National Aeronautics and Space Administration under Grant NSG-2265, and the U.S. Office of Naval Research under the Joint Services Electronics Program Contract N00014-75-C-0648.

There is a large class of processes whose statistical properties are easy to study and which are governed by linear stochastic differential equations of the form

$$dx = Axdt + \sum_{i=1}^{r} B_i xdw_i + \sum_{i=1}^{s} C_i xdN_i + \sum_{i=1}^{r} d_i dw_i + \sum_{i=1}^{s} e_i dN_i$$

This is the class of models we will be concerned with here.

The remainder of the paper is organized as follows. The next two sections are devoted to a discussion of examples of processes which are solutions of stochastic equations involving dw's and dN's, respectively. The fourth section discusses some combined effects. Estimation and the concepts of observability appropriate in this context are then discussed and finally we give a number of references to the literature.

II. BILINEAR EQUATIONS WITH WIENER PROCESSES. The first feature of a model of the form

$$dx = -\alpha xdt + \beta xdw + \gamma dw$$

is that all moments which exist at $t = 0$ exist for $0 \leq t < \infty$ and can be solved for by solving a finite set of linear equations. The Itô differentiation rule is the key. If x satisfies a scalar equation of the above form then its pth power satisfies

$$dx^p = -\alpha px^p dt + \beta px^p dw + \gamma px^{p-1} dw + \frac{1}{2}\, p(p-1)x^{p-2}(\beta^2 x^2+\gamma^2)dt$$

In view of the way the Itô integral is defined we obtain, on taking expectations,

$$\frac{d}{dt}\, \mathscr{E} x^p = (-\alpha p+ \frac{1}{2}p(p-1)\beta^2)\, \mathscr{E} x^p + \frac{1}{2}p(p-1)\, \mathscr{E} x^{(p-2)}\gamma^2$$

These linear equations give the pth moment in terms of the (p-2)th moment, etc. These calculations make it clear that for scalar systems sufficiently high order moments will be growing with time regardless of how stable the deterministic part of the equation is. Incidentally, models of the form

$$dx = Axdt + \sum_{i=1}^{r} B_i xdw_i + \sum_{i=1}^{r} g_i dw_i$$

are only superficially more general than those with no additive noise term,

$$dx = Axdt + \sum_{i=1}^{r} B_i xdw_i$$

because by introducing in place of x

$$\tilde{x} = \begin{bmatrix} 1 \\ x \end{bmatrix}$$

we can write

$$d\tilde{x} = \begin{bmatrix} 0 & 0 \\ 0 & A \end{bmatrix} \tilde{x}dt + \sum_{i=1}^{r} \begin{bmatrix} 0 & 0 \\ g_i & B_i \end{bmatrix} \tilde{x}dw_i$$

and in this way remove the need for additive terms.

There is a large class of processes whose statistical properties are easy to study and which are governed by linear stochastic differential equations of the form

$$dx = Axdt + \sum_{i=1}^{r} B_i x dw_i + \sum_{i=1}^{s} C_i x dN_i + \sum_{i=1}^{r} d_i dw_i + \sum_{i=1}^{s} e_i dN_i$$

This is the class of models we will be concerned with here.

The remainder of the paper is organized as follows. The next two sections are devoted to a discussion of examples of processes which are solutions of stochastic equations involving dw's and dN's, respectively. The fourth section discusses some combined effects. Estimation and the concepts of observability appropriate in this context are then discussed and finally we give a number of references to the literature.

II. <u>BILINEAR EQUATIONS WITH WIENER PROCESSES</u>. The first feature of a model of the form

$$dx = -\alpha x dt + \beta x dw + \gamma dw$$

is that all moments which exist at $t = 0$ exist for $0 \leq t < \infty$ and can be solved for by solving a finite set of linear equations. The Itô differentiation rule is the key. If x satisfies a scalar equation of the above form then its pth power satisfies

$$dx^p = -\alpha p x^p dt + \beta p x^p dw + \gamma p x^{p-1} dw + \frac{1}{2} p(p-1) x^{p-2}(\beta^2 x^2 + \gamma^2) dt$$

In view of the way the Itô integral is defined we obtain, on taking expectations,

$$\frac{d}{dt} \mathscr{E} x^p = (-\alpha p + \frac{1}{2} p(p-1)\beta^2) \mathscr{E} x^p + \frac{1}{2} p(p-1) \mathscr{E} x^{(p-2)} \gamma^2$$

These linear equations give the pth moment in terms of the (p-2)th moment, etc. These calculations make it clear that for scalar systems sufficiently high order moments will be growing with time regardless of how stable the deterministic part of the equation is. Incidentally, models of the form

$$dx = Axdt + \sum_{i=1}^{r} B_i x dw_i + \sum_{i=1}^{r} g_i dw_i$$

are only superficially more general than those with no additive noise term,

$$dx = Axdt + \sum_{i=1}^{r} B_i x dw_i$$

because by introducing in place of x

$$\tilde{x} = \begin{bmatrix} 1 \\ x \end{bmatrix}$$

we can write

$$d\tilde{x} = \begin{bmatrix} 0 & 0 \\ 0 & A \end{bmatrix} \tilde{x} dt + \sum_{i=1}^{r} \begin{bmatrix} 0 & 0 \\ g_i & B_i \end{bmatrix} \tilde{x} dw_i$$

and in this way remove the need for additive terms.

Higher dimensional bilinear models need not have higher order moments which are unstable. Consider a two dimensional model

$$d\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1/2 & 1 \\ -1 & -1/2 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} dt + \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} dw_1 + \begin{bmatrix} 0 \\ \sqrt{2} \end{bmatrix} dw_2$$

In this case a complete analysis of the moments involves the study of equations for $x^{[p]}$ which is defined by

$$x^{[p]} = \begin{bmatrix} x_1^p \\ \alpha_1 x_1^{p-1} x_2 \\ \alpha_2 x_1^{p-2} x_2^2 \\ \vdots \\ x_2^p \end{bmatrix}$$

with the α_i chosen to be positive and such that

$$(x_1^2+x_2^2) = (x_1^p)^2 + \alpha_1^2(x_1^{p-1}x_2)^2 + \alpha_2^2(x_1^{p-2}x_2^2)^2 + \ldots + (x_2^p)^2$$

Using the Itô rule again we get a linear stochastic differential equation for $x^{[p]}$ of the form

$$dx^{[p]} = \tilde{A}x^{[p]}dt + \tilde{B}x^{[p]}dw_1 + \tilde{g}x^{[p-2]}dw_2 + fdt$$

It turns out that because of the skew symmetry of the matrix multiplying xdw_1 the multiplicative noise term acts to decrease the higher order moments see [2]. For this example it turns out that in steady state x has a Gaussian distribution with variance I and mean zero. This can be seen, by looking at the moments of all order or by looking at the Fokker-Plank equation. However the process x is not Gauss-Markov.

One of the properties of the higher order correlations of stationary Gauss-Markov processes is the simple relationship between the rates of decay of the various correlations. For example, if x is Gauss-Markov and

$$\mathcal{E}x(t)x(t+\tau) = e^{-\tau} \quad ; \qquad \tau > 0$$

then

$$\mathcal{E}x(t)x(t+\tau)x(t+\tau+\sigma)x(t+\tau+\sigma+\rho) = e^{-(\tau+\sigma+\rho)}; \qquad \tau,\sigma,\rho > 0$$

For the above process the fourth order correlations follow a different, but still easily calculated, pattern.

The key to calculation of these correlations is to note that for

$$dx = Axdt + \sum_{i=1}^{n} B_i x dw_i + \sum_{i=1}^{r} g_i dw_i$$

We can take expectations and get

$$\frac{d}{dt}\mathscr{E}x(t) = A\mathscr{E}x(t)$$

This means that

$$\mathscr{E}x(t)x'(t+\tau) = \Sigma(t)e^{A\tau} \; ; \qquad \tau > 0$$

where

$$\Sigma(t) = \mathscr{E}x(t)x'(t)$$

Since $x^{[p]}$ satisfies a bilinear stochastic equation we can use similar ideas to calculate its correlations.

III. BILINEAR SYSTEMS WITH POISSON COUNTERS. Let N be a standard Poisson counter of rate λ. The process x generated by

$$dx = -2xdN$$

evolves as follows. It is constant until N jumps. Because our stochastic integrals are Itô integrals, at a jump x changes from its present value x^- to $x^- - 2x^- = -x^-$. Thus x flips back and forth from x(0) to -x(0) to x(0), etc. If x(0) = 1 it is a random telegraph wave with amplitudes ±1 and rate λ. We can use this representation of the random telegraph wave together with the Ito rule to compute various statistical properties in a totally straightforward way. Because $\mathscr{E}(dN-\lambda dt)\phi(x) = 0$ for reasonable functions ϕ we have

$$\begin{aligned}\frac{d}{dt}\mathscr{E}x &= -2\mathscr{E}x(dN-\lambda dt)-2\lambda\mathscr{E}x \\ &= -2\lambda\mathscr{E}x\end{aligned}$$

In using the Ito calculus with Poisson processes it is most convenient to calculate the effect on $\phi(x)$ of a jump in x and then write down the appropriate differential equation. For example, if x satisfies the above equation then

$$dx^2 = 0$$

because a jump in x of the type $x \mapsto -x$ does not change x^2. Thus we see, for example, that in steady state

$$\mathscr{E}x(t)x(t+\tau) = e^{-2\lambda\tau}$$

A finite-state continuous time jump process is one which takes on a finite set of values, say $x_1, x_2, \ldots x_r$ and jumps between these values according to a probability law of the form

$$\begin{bmatrix}\dot{p}_1\\ \dot{p}_2\\ \vdots\\ \dot{p}_r\end{bmatrix} = \begin{bmatrix}a_{11} & a_{12} & \cdots & a_{1r}\\ a_{21} & a_{22} & & a_{2r}\\ \cdots & \cdots & \cdots & \cdots\\ a_{r1} & a_{r2} & & a_{rr}\end{bmatrix}\begin{bmatrix}p_1\\ p_2\\ \vdots\\ p_r\end{bmatrix}$$

These models are very useful in queuing theory etc. We can associate a numerically valued stochastic process with a finite state continuous time jump process in much the same way as we did with the random telegraph wave above. We just assign a numerical value to each state. The processes one can get in this way are in some senses quite general. In [3] we showed that any stationary process can be matched, with arbitrary precision, by such a process insofar as the mean and covariance are concerned. We also know [1] that finite state continuous time jump processes can be modeled by bilinear stochastic differential equations of the form

$$dx = A_i x dN_i$$

The idea here is as follows. Let x take on only the values $(e_1, e_2, \ldots, e_r)$ where

$$e_i = (0\ldots0,\underset{\text{ith position}}{1},0,\ldots0)'$$

That is we code the states as unit vectors in an r-dimensional space. If each of the A_i are matrices of the form

$$\begin{array}{ccc} & \begin{array}{ccccc} & & & \downarrow\text{ith column} & \end{array} & \\ E_{ij} = & \begin{bmatrix} 0 & 0 & \cdots & & 0 \\ 0 & 0 & \cdots & & 0 \\ & & & 1 & \\ 0 & 0 & \cdots & & 0 \end{bmatrix} & \begin{array}{c} \\ \\ \longleftarrow \text{jth row} \\ \\ \end{array} \end{array}$$

Then x will jump from one unit vector to another with a probability law which is determined by the rates of the counters N_i. This representation of a finite state process is perfectly general although it will typically require r(r-1) independent counting processes.

IV. COMBINED MODELS. One extremely unpleasant feature of Gauss-Markov models is that bilinear models of the form

$$dx = Ax + \eta Bx + g\gamma$$

with η Gauss-Markov but not white are virtually intractable. The variance of x cannot be calculated, etc. Gauss-Markov multiplicative noise processes which are not white lead to trouble! The situation for finite state processes is very much nicer. This fact, coupled with the natural suitability of finite state models in reliability modeling make it worthwhile to consider this in some detail.

We examine the scalar equation

$$dx = -3xdt + mxdt + dw$$

where m(t) is w random telegraph wave of rate λ. For example, $m(t) = 1$ might correspond to a failed state and $m = -1$ a working state. If we wish to determine the state of m from observations on x we need to know the statistical properties of x. All such calculations are made routine by writing

$$dm = -2mdN$$

and considering the triple (m,x,mx). We get (remember $m^2 = 1$)

$$d\begin{bmatrix} m \\ x \\ mx \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 3dt & dt \\ 0 & dt & -3dt \end{bmatrix}\begin{bmatrix} m \\ x \\ mx \end{bmatrix} + \begin{bmatrix} 2dN & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}\begin{bmatrix} m \\ x \\ mx \end{bmatrix} + \begin{bmatrix} 0 \\ dw \\ 0 \end{bmatrix}$$

We have

$$\frac{d}{dt}\mathscr{E}\begin{bmatrix} x \\ mx \end{bmatrix} = \begin{bmatrix} -3 & 1 \\ 1 & -2\lambda-3 \end{bmatrix}\mathscr{E}\begin{bmatrix} x \\ mx \end{bmatrix}$$

and we can compute the variance of x together with that of mx by examining the differential equation for

$$\begin{bmatrix} x \\ mx \end{bmatrix}[x,mx] = \begin{bmatrix} x^2 & mx^2 \\ mx^2 & x^2 \end{bmatrix}$$

V. ESTIMATION PROBLEM*. There are several features of bilinear equations which make them interesting examples on which to attempt nonlinear estimation. In the first place the optimal linear filter is easily constructed. Secondly one can show rather easily in most cases that the best linear estimate is not the best estimate. Finally, we will show how to use tensor methods to construct nonlinear state estimators whose performance improves along with the degree of nonlinearity.

To start with consider the process and observation

$$dx = -(1+\frac{1}{2}\alpha^2)xdt+\alpha xdw_1+dw_2$$

$$dy = xdt+dw_3$$

The steady state variance and the covariance of the x process is given by

$$\mathscr{E}x(t)x(t+\tau) = \frac{1}{2}e-|\tau|$$

The best steady state linear filter is the same as that associated with the Gauss-Markov model

$$d\tilde{x} = -\tilde{x}dt+dw_2; \qquad dy = \tilde{x}dt+dw_3$$

and has mean square error $k = \sqrt{2} -1$. The best linear estimate is generated by

$$dz = -zdt + (\sqrt{2} -1)(xdt+dw_3-zdt)$$

*Some interesting work on the bilinear estimation problem has been done recently at MIT by S.K. Mitter and D. Ocone. However this is as yet unpublished and we will not go into it here.

The optimal linear estimator is independent of α and, of course, for $\alpha = 0$ it generates the optimal estimate. For $\alpha \neq 0$ this filter is not optimal because the error $(x-z)$ is correlated with z^p for p odd. That means, of course that we can reduce the mean square error by replacing the estimate z by f(z) for a suitable function f. Calculations of this type are not too difficult to make; the following is illustrative. The differential equation for x and z are

$$d\begin{bmatrix} x \\ z \end{bmatrix} = \begin{bmatrix} -(1+\frac{1}{2}\alpha^2)dt+\alpha dw_1 & 0 \\ (\sqrt{2}-1)dt & -\sqrt{2} \end{bmatrix}\begin{bmatrix} x_1 \\ z \end{bmatrix} + \begin{bmatrix} dw_2 \\ (\sqrt{2}-1)dw_3 \end{bmatrix}$$

By using Itô's rule we can get a linear equation for

$$\begin{bmatrix} x \\ z \end{bmatrix}^{[3]} = \begin{bmatrix} x^{[3]} \\ \sqrt{3}\, x^{[3]}z \\ \sqrt{3}\, xz^2 \\ z^3 \end{bmatrix}$$

which, in turn, contains all the data needed to find the values of α and β such that

$$\eta = \mathscr{E}(x-\alpha z-\beta z^3)^2$$

is minimized.

It should be noted that since

$$dx^p = -p(1+\frac{1}{2}\alpha^2)x^p+p\alpha x^p dw+x^{p-1}dw+\frac{1}{2}p(p-1)(\alpha^2x^p+x^{(p-2)})dt$$

the pth moment of x exists only if $(1+\frac{1}{2}\alpha^2) > \frac{1}{2}(p-1)\alpha^2$. That is, as α gets larger fewer moments of x (and hence z) exist. Apparently moment techniques are not the appropriate way to treat this problem for α large. On the other hand, it is exactly when α is large that we expect to gain the most improvement over the performance of the linear filter.

To begin a general discussion we recall a few facts about the standard linear state estimation problem. Consider

$$dx = Axdt + Bdw; \quad dy = Cxdt + Dd\nu$$

where (w,ν) is a vector of independent standard Wiener processes. If we wish to estimate x by z we set

$$dz = Azdt + L(Cxdt+Dd\nu-Cz)$$

This gives an error $e = (x-z)$ which satisfies

$$de = (A+LC)edt + Bdw + LDdw$$

and the variance for the error of $P(t,t) = \mathcal{E}e(t)e'(t)$ which satisfies

$$\frac{d}{dt}P(t,t) = (A+LC)P(t,t)+P(t,t)(A+LC)'+BB'+LDD'L'$$

setting $L = KC'$ we note an analogy with the linear quadratic optimal control problem and observe that if

$$\dot{K} = AK+KA'+KC(DD')C'K'+BB'$$

Then we minimize $P(t,t)$. The condition of observability of the pair $(A,(DD')^{1/2}C')$ insures finite error variance even if the state equation is unstable.

Now consider the stochastic process generated by

$$dx = Axdt + \sum_{i=1}^{m} B_i x dw_i$$

with the observation $dy = Cxdt + \sum_{i=1}^{p} D_i x d\nu_i$. The minimum variance linear filter is defined by

$$dz = Azdt + L(Cx-Cz + \sum_{i=1}^{p} LD_i x d\nu_i)$$

then the error x-z satisfies

$$de = (A+LC)edt + \sum_{i=1}^{m} B_i x dw_i + \sum_{i=1}^{p} LD_i x d\nu_i; \quad L = \Sigma_{ee}C$$

and the variance Σ_{ee} satisfies

$$\dot{\Sigma}_{ee} = (A+LC)\Sigma_{ee}+\Sigma_{ee}(A+LC)' + \sum_{i=1}^{m} B_i\Sigma_{xx}B_i' + \sum_{i=1}^{p} LD_i\Sigma_{xx}D_i'L'$$

where Σ_{xx} is the a priori expected value of $x(t)x'(t)$. That is to say, the minimum variance linear filter is the same as that of the linear system

$$d\eta = A\eta dt + Gdw; \qquad dy = (\Sigma_{xx})^{1/2}D'd\eta + C\eta dt$$

where $GG' = \sum_{i=1}^{m} B_i\Sigma_{xx}B_i$.

One aspect of the bilinear estimation problem which sets it apart from others is the fact that one may use a simple algebraic trick to improve the accuracy of the best linear filter. The idea is that if the original model is

$$dx = Axdt + \sum_{i=1}^{r} B_i x dw_i; \qquad dy = Cxdt + dw$$

then there is, for each p, a related system in $x^{[p]}$

$$dx^{[p]} = \tilde{A}x^{[p]} + \sum_{i=1}^{r} \tilde{B}_i x^{[p]} dw_i$$

and, of course, if we observe dy we can regard dy^p as being observed as well.

The main point is that the best linear estimate for the system satisfied by

$$x_{[p]} = \begin{bmatrix} x \\ x^{[2]} \\ \vdots \\ x^{[p]} \end{bmatrix}$$

with observation $dy, dy^2, \ldots dy^p$ generally provides a better estimate for x than does the best linear estimate based on $x_{[p-1]}$. Of course we must make sure that the moments introduced in this way actually exist.

A second interesting aspect of the estimation problem is that it may happen that states are "unobservable" in the original system and only become "observable" in the $x_{[p]}$ setup. We illustrate this somewhat vague remark as follows. Consider

$$dr = -rdt + dw_1$$

$$dx = -xdt + rdw_2; \quad dy = xdt + dw_3$$

In steady state the best linear estimator will estimate r as a constant, in fact zero. However we can get an estimate for r^2 which is better than the a priori value of 1/2 by considering the system

$$dr^2 = -2dr^2dt+2pdw_1+dt$$

$$dx^2 = -2x^2dt+2rxdw_2+r^2dt; \quad dy^2=(x+w_3)(xdt+dw)+dt$$

$$dxr = -2xrdt+xdw_1+r^2dw_2$$

At this level the interaction between x and r is such that a linearization results in a system in which r^2 is observable.

As a final example consider the stochastic differential equation

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} -dt & dw_1 \\ -dw_1 & -\frac{1}{2}dt \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} dw_2 \\ 0 \end{bmatrix}; \quad dy = x_1dt+dw_3 \qquad (*)$$

where w_1, w_2, and w_3 are all standard Wiener processes. The steady state density for x is Gaussian in this case with variance I. One sees without too much trouble that in steady state

$$\mathcal{E}x(t)x'(t+\tau)= \begin{bmatrix} e^{-|\tau|} & 0 \\ 0 & e^{-\frac{1}{2}|\tau|} \end{bmatrix}$$

and that

$$F(\tau) = \begin{bmatrix} x_1^2(t)x_1^2(\tau) & x_1^2(t)x_2^2(\tau) \\ x_1^2(\tau)x_2^2(t) & x_2^2(t)x_2^2(\tau) \end{bmatrix}' = \begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix} \exp \begin{bmatrix} -2 & 1 \\ 1 & -1 \end{bmatrix} \tau$$

Instead of being

$$F_g(\tau) = \begin{bmatrix} 3e^{-2|\tau|} & 0 \\ 0 & 3e^{-|\tau|} \end{bmatrix}$$

as it would be for a Gauss-Markov process.

This system is "second order unobservable" in the sense that the covariance for x_1 is that of a first order system. We see rather easily that the best steady state linear filter for x_1 is given by

$$d\hat{x} = -\hat{x}dt + (\sqrt{3}-1)(x_1dt+dw_3-\hat{x}dt)$$

where the coefficient $(\sqrt{3}-1)$ comes from solving the error variance equation

$$\dot{\sigma} = -2\sigma + 2 - \sigma^2$$

for its steady state value. Note the steady state error has a variance of $\sqrt{3}-1 = .73 \ldots$ and $\tilde{x}$ has a steady state variance of $2-\sqrt{3} = .27\ldots$

Putting together equation (*) and the best linear filter we have

$$\begin{bmatrix} x_1 \\ x_2 \\ x \end{bmatrix} = \begin{bmatrix} -dt & dw_1 & 0 \\ -dw_1 & -\frac{1}{2}dt & 0 \\ (\sqrt{3}-1) & 0 & -\sqrt{3} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x \end{bmatrix} + \begin{bmatrix} dw_2 \\ 0 \\ (\sqrt{3}-1)dw_3 \end{bmatrix}$$

and thus

$$\frac{d}{dt}\begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{bmatrix} = \begin{bmatrix} -1 & 0 & 0 \\ 0 & -\frac{1}{2} & 0 \\ (\sqrt{3}-1) & 0 & -\sqrt{3} \end{bmatrix}\begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{bmatrix} + \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{bmatrix}$$

$$\times \begin{bmatrix} -1 & 0 & \sqrt{3}-1 \\ 0 & -\frac{1}{2} & 0 \\ 0 & 0 & -\sqrt{3} \end{bmatrix} + \begin{bmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}\begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{bmatrix}\begin{bmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & (\sqrt{3}-1)^2 \end{bmatrix}$$

From this we deduce that in steady state

$$\mathcal{E}\begin{bmatrix} x_1(t) \\ x_2(t) \\ \hat{x}(t) \end{bmatrix} [x_1(t+\tau) \quad x_2(t+\tau) \quad x_3(t+\tau)]$$

$$= \begin{bmatrix} 1 & 0 & 2-\sqrt{3} & e^{-\tau} \\ 0 & 1 & 0 & 0 \\ 2-\sqrt{3} & 0 & (2-\sqrt{3}) & 0 \end{bmatrix} \begin{bmatrix} 0 & \frac{1}{\sqrt{3}-1}(e^{-\tau}-e^{-\sqrt{3}\tau}) \\ e^{-1/2\tau} & 0 \\ 0 & e^{-\sqrt{3}\tau} \end{bmatrix}$$

By passing to an $x_{[p]}$ version of these equations we get still different kinds of nonlinear filters whose performance will improve on the linear filter. Obviously a large number of variations on this theme are possible.

VI. REFERENCES

1. R.W. Brockett and G. Blankenship, "A Representation Theorem for Linear Differential Equations with Markovian Coefficients," 1977 Allerton Conf.

2. R.W. Brockett, "Parametrically Stochastic Linear Systems," in Stochastic Systems: Modeling, Identification, and Optimization, (Studies in Nonlinear Programming, Vol. 5), pp. 8-21, North Holland Publishers, 1976.

3. R.W. Brockett, "Stationary Covariance Generation with Fintie State Markov Processes," 1977 Joint Automatic Control Conference.

4. R.W. Brockett, "Stochastic Bilinear Models," 1977 Joint Automatic Control Conference.

5. R.W. Brockett, "Lie Algebras and Lie Groups in Control Theory," Geometric Methods in System Theory, Reidel Publishing Co., Dordrecht, The Netherlands, (D.Q. Mayne and R.W. Brockett, eds.), 1973, pp. 43-82.

6. R.W. Brockett, "Lie Theory and Control Systems Defined on Spheres," SIAM J. on Applied Mathematics, Vol. 25, No. 2, Sept. 1973, pp. 213-225.

7. R.W. Brockett, "Volterra Series and Geometric Control Theory," Automatica, Vol. 12, No. 2, March 1976, pp. 167-176.

8. J. Morrison and J. McKenna, "Analysis of Some Stochastic Ordinary Differential Equation," in Stochastic Differential Equations, SIAM-AMS Proc., Vol. 6, 1973, pp. 97-161.

9. E. Wong, Stochastic Processes in Information and Dynamical Systems, McGraw-Hill, N.Y. 1971.

10. W. Wonham, "Random Differential Equations in Control Theory," in Probabilistic Methods in Applied Mathematics, Vol. 2, A. Bharucha-Reid, ed., Academic Press, N.Y. 1970, pp. 131-215.

11. W.M. Wonham, "Some Applications of Stochastic Differential Equations to Optimal Nonlinear Filtering," J. SIAM, Series A: Control, 2(3), 1965.

12. W.M. Wonham, "Optimal Stationary Control of a Linear System with State-Dependent Noise," SIAM J. Control, Vol. 5, 1967, pp. 486-500.

HYPERBOLIC CONSERVATION LAWS

Ronald J. DiPerna
Mathematics Research Center
Madison, Wisconsin 53706

ABSTRACT. We shall discuss some basic results in the theory of conservation laws and comment on their connection with the numerical computation of the solution.

I. INTRODUCTION.

In this lecture we shall be concerned with the initial-value problem for systems of conservation laws,

$$\frac{\partial}{\partial t} U + \sum_{j=1}^{m} \frac{\partial}{\partial x_j} F^j(U) = 0 .$$

Here the solution $U = U(x_1, x_2, \dots x_m, t)$ takes on values in R^n and F^j is a smooth nonlinear mapping from R^n to R^n. Equations of this form arise in continuum mechanics. The equations of gas dynamics and thermoelasticity form systems of five equations; the components of U represent the densities of mass, linear momentum (3), and total energy while the equations express the physical laws for the conservation of the corresponding five quantities, cf. [3, 9]. Other examples are provided by magneto-fluid dynamics, elasticity and the theory of shallow water waves. Equations of this type present a variety of problems of mathematical and engineering interest. Unfortunately, a rigorous mathematical theory is yet to be developed for systems in several space dimensions. However, there has been a great deal of progress on systems in one space dimension and we shall discuss a couple of basic results in this direction.

II. SYSTEMS IN ONE SPACE DIMENSION.

It is natural to ask: what type of solutions will a system of conservation laws generate? A preview can be obtained by considering a scalar equation in one space dimension,

$$\frac{\partial}{\partial t} u + \frac{\partial}{\partial x} f(u) = 0 \quad , \quad -\infty < x < \infty . \tag{1}$$

Here f is a smooth nonlinear mapping from R to R. The classical theory of conservation laws guarantees that if the initial data are smooth then there exists a smooth solution defined for a small interval of time. More precisely, if $u_0(x) \varepsilon C^1$ then there exists a C^1 solution $u(x,t)$ which is defined in some strip $0 \leq t < T$ and which takes on the initial data $u_0(x)$ at $t = 0$; the length of existence T depends only on the C^1-norm of the data. Presently, we shall see how to construct C^1 solutions. But first let us assume that we are given a C^1 solution defined on some strip $0 \le t < T$ and examine how the values of u propagate.

If u is a C^1 solution, we may carry out the differentiation in equation (1) with respect to x,

$$\frac{\partial}{\partial t}u + f'(u)\frac{\partial}{\partial x}u = 0\,, \tag{2}$$

and introduce characteristic curves $(x(t),t)$ which are defined by

$$\frac{d}{dt}x(t) = f'\{u(x(t),t)\}\,,\ x(0) = x_0\,. \tag{3}$$

Here $x(t)$ denotes the position of the characteristic as a function of time and $\frac{d}{dt}x(t)$ its speed of propagation. Since we are assuming that the solution $u = u(x,t)$ is known, the right hand side of (3), i.e. f composed with u, is a known function of x and t and the standard theory for ordinary differential equations guarantees the existence of a characteristic curve through each point x_0 on the initial line $t = 0$.

We first observe that the solution u is constant along characteristic curves. Indeed, the restriction of u to a characteristic curve $(x(t),t)$, i.e.

$$u(x(t),t)\,,$$

satisfies

$$\frac{d}{dt}u(x(t),t) = \frac{\partial}{\partial x}u(x(t),t)\frac{d}{dt}x(t) + \frac{\partial}{\partial t}u(x(t),t)\,. \tag{4}$$

Using the definition (3) of characteristic curves and the equation (2) we obtain

$$\frac{d}{dt}u(x(t),t) = \frac{\partial}{\partial x}u(x(t),t)f'\{u(x(t),t)\} + \frac{\partial}{\partial t}u(x(t),t) = 0.$$

Thus, if u is a C^1 solution we may interpret the equation (2) as the statement that the directional derivative of u in the direction prescribed by the characteristic field vanishes.

Since u is constant along characteristic curves, so is the speed of propagation $f'(u)$: characteristic curves are simply straight lines. We may now use these two facts to establish local existence of C^1 solutions. Suppose now that we are given C^1 initial data $u_0(x)$. Let us define a function $u = u(x,t)$ by the requirement that u equal the value $u_0(x_0)$ along the line

$$x = x_0 + f'(u_0(x_0))t\,,$$

i.e. along the characteristic through $(x_0,0)$. It is not difficult to show using the implicit function theorem that there exists a time T such that the above function u is a C^1 solution for $0 \leq t < T$ and that T depends only on the C^1 norm of the initial data $u_0(x)$.

The method of characteristics answers the question of local existence and points to the main source of mathematical difficulty associated with the equation, the focusing of waves. In general, different characteristic curves carrying different values of the solution intersect. In the neighborhood of such a point there can not exist a classical C^1 solution. On the other hand, for problems of physical origin one certainly expects to have a·globally defined solution. In the flow of gas, for example, one sees experimentally that the focusing of sound waves leads to the development of shock waves, i.e. abrupt changes in the physical variables which occur over a distance of a few mean free paths of the molecules. In the hyperbolic (or inviscid) theory of gas dynamics, such flows are modeled by discontinuous functions and the mathematical theory of conservation laws is developed in the framework of weak solutions, i.e. solutions which satisfy the equations in the sense of distributions.

Global existence of weak solutions for systems of conservation laws in one space dimension was established by Glimm [7]. Glimm considered strictly hyperbolic systems of the form

$$\frac{\partial}{\partial t} U + \frac{\partial}{\partial x} F(U) = 0 \quad , \quad -\infty < x < \infty . \tag{5}$$

Here $U = U(x,t) \varepsilon R^n$ and F is a smooth nonlinear mapping from R^n to R^n whose Jacobian matrix has n real and distinct eigenvalues. We note that the hypothesis of strict hyperbolicity is satisfied by many of the conservative systems of physical interest in one space dimension: gas dynamics, magneto-fluid dynamics, shallow water waves and in certain cases elasticity. The theorem of Glimm may be stated as follows: if TVU_0 is sufficiently small then there exists a globally defined weak solution of (5) which assumes the initial data $U_0(x)$ at $t = 0$ and which satisfies,

$$TVU(\cdot,t) \leq \text{const. } TVU_0 \tag{6}$$

where the constant depends only on the nonlinear term F.

Glimm's result contains a method for the construction of the solution which has recently been implemented for the purposes of numerical calculation, cf. [2]. The method has also been extended to the construction of solutions with initial data having large total variation, [1,4,5,8,10,12,13].

Before commenting on estimate (6) in connection with the numerical computation of the solution we shall briefly discuss functions of bounded variation in one variable and the physical interpretation of (6). Suppose that g is an arbitrary function mapping R to R. Consider any finite set of pts $\{x_j : j = 1,2...n\}$ satisfying

$$x_1 < x_2 < \dots < x_n ,$$

together with the sum of all the associated increments,

$$\sum_{k=1}^{n-1} |g(x_{k+1}) - g(x_k)| .$$

The total variation of g is defined as the supremum of the sum of all increments associated with all possible finite ordered partitions of R:

$$TVg = \sup\left[\sum_{k=1}^{n-1} |g(x_{k+1}) - g(x_k)| : \text{all } \{x_j : y = 1,2,\dots n\}\right] .$$

If TVg is finite the function g is said to have finite total variation. A vector-valued function is said to have finite total variable if all of its components do. As a simple example we mention the class of piecewise constant functions with a finite number of jumps. For such functions the total variation equals the sum of all the jumps. Secondly, we mention the class of smooth functions whose first derivative is integrable. For such functions g we have

$$TVg = \int_{-\infty}^{\infty} |g'(x)|dx .$$

We refer the reader to [11] for the theory of functions of bounded variation of one variable and to [6, 15] for the theory functions of bounded variation of several variables.

In the context of conservation laws the total variation of the solution U(x,t) with respect to the space variable x at a fixed time t represents the total magnitude of all waves in the solution at time t, all shock waves, rarefaction waves, compression waves, etc. The estimate (6) may be interpreted as stating that the total magnitude of a waves in the solution at time is bounded (uniformly in t) by some multiple of the total magnitude of all waves in the initial data.

From the point of view of numerical analysis it would be of interest to determine whether or not the difference schemes in current usage are stable in the total variation norm [14]. A difference scheme is said to be stable in the total variation norm if

$$TVU_h(\cdot,t) \le \text{const.},$$

where $U_h(x,t)$ denotes the piecewise constant approximate solutions generated by the scheme using mesh length h and where the constant is independent of h and t. The constant would of course depend upon F and total variation of the initial data $U_0(x)$. It would be of interest, in particular, to test various difference schemes for stability in the total variation norm in the case of Riemann initial data, i.e.

$$U(x,0) = \begin{cases} U^+ & \text{if } x > 0 \\ U^- & \text{if } x < 0 \end{cases}$$

where U^+ and U^- are constant states.

Lastly, we note that the total variation norm is one of the most simple and natural norms for systems of conservation laws which is sufficiently strong to guarantee that stability in the norm implies convergence of the scheme. It is presently an open problem to prove the convergence of difference schemes for (5).

References

1. Bakhvarov, N., On the existence of regular solutions in the large for quasilinear hyperbolic systems, Zhur. Vychisl. Mat. i Mathemat. Fiz., 10, (1970), 969-980.

2. Chorin, J. A., Random choice solution of hyperbolic systems, to appear in J. Comp. Phys.

3. Courant, R. & K. O. Friedrichs, "Supersonic Flow and Shock Waves", New York: Interscience Publishers, Inc. 1948.

4. DiPerna, R. J., Global solutions to a class of nonlinear hyperbolic systems of equations, Comm. Pure Appl. Math., 26 (1973), 1-28.

5. DiPerna, R. J., Existence in the large for nonlinear hyperbolic conservation laws, Arch. Rat. Mech. Anal., 52 (1973) 244-257.

6. Federer, H., "Geometric Measure Theory", New York: Springer 1969.

7. Glimm, J., Solutions in the large for nonlinear hyperbolic systems of equations, Comm. Pure Appl. Math., 18 (1965), 697-715.

8. Greenberg, J. M., The Cauchy problem for the quasilinear wave equation, unpublished.

9. Lax, P. D., Hyperbolic systems of conservation laws, II, Comm. Pure Appl. Math.,10 (1957), 537-566.

10. Liu, T.-P., Solutions in the large for the equations of non-isentropic gas dynamics, Indiana Univ. Math. J., 26 (1977), 147-177.

11. Natanson, I. P., "Theory of Functions of a Real Variable", New York: Frederick Ungar 1955.

12. Nishida, T., Global solutions for an initial boundary value problem of a quasilinear hyperbolic system, Proc. Japan Acad., 44 (1968), 642-646.

13. Nishida, T. & J. A. Smoller, Solutions in the large for some nonlinear hyperbolic conservation laws, Comm. Pure Appl. Math., (1973), 183-200.

14. Sod, G. A., A survey of numerical methods for compressible fluids, ERDA Math. and Computer Lab. Report, Courant Institute, N.Y.U. 1977.

15. Vol'pert, A. I., The spaces BV and quasilinear equations, Math. USSR Sb., 2 (1967), 257-267.

ANALYSIS OF A STOCHASTIC REYNOLDS EQUATION AND RELATED PROBLEMS*

P. L. Chow**

Department of Mathematics, Wayne State University
Detroit, Michigan 48202

E. A. Saibel

Engineering Sciences Division, U. S. Army Research Office
Research Triangle Park, North Carolina 27709

ABSTRACT. This work is mainly concerned with the analysis of a stockastic Reynolds equation in the hydrodynamic theory of lubrication. A differential equation for the mean pressure distribution is derived rigoriously in the asymptotic limit by invoking an ergodic theorem. Also two upper bounds to the absolute mean and the root-mean-square deviations of a normalized load carrying capacity from the smooth case is obtained for a general one-dimensional problem. Finally some related problems are discussed briefly.

I. INTRODUCTION.

In the present paper, we shall first summarize two main results contained in our recent paper [1], to which the reader is referred for a detailed presentation. The first result is concerned with the derivation of an valid, averaged Reynolds equation under appropriate conditions. Broadly stated, the sufficient conditions require that the roughness parameters for the two surfaces be characterized by weakly dependent random functions of the longitudinal variable, and that the length of the bearing be large relative to the correlation length of the roughness parameters. As our second result, two upper bounds for the absolute mean and the root-mean-square deviations of

*Presented at 24th Conference of Army Mathematicians, 2 June 1978, Charlottesville, Virginia.

**The work of this author was supported by the U. S. Army Research Office Grant DAAG29-78-G-0042.

a normalized load carrying capacity from a smooth case are obtained for a general one-dimensional problem, valid for an arbitrary probability distribution. The result shows a critical dependence of the upper bounds on the correlation of the roughness. These results are presented in section II.

In the last section, we shall give a general discussion of some related questions, the two dimensional problem and its connection with other random boundary problems in a different physical context.

II. ANALYSIS OF A STOCHASTIC REYNOLDS EQUATION.

For a wide slider bearing of length L, the Reynolds equation in one dimension reads

$$\frac{d}{dx}\left(H^3 \frac{dp}{dx}\right) = \Lambda\frac{d}{dx}\left[h + v(\delta_1 - \delta_2)\right] , \qquad (1)$$

$$p(o) = p(L) = o . \qquad (2)$$

where $p = p(x)$ is the pressure,

$H = H(x) = h(x) + \delta_1(x) + \delta_2(x)$ is the film thickness,

$h(x) = \langle H(x) \rangle$ is the average (or mean) film thickness,

$\delta_1(x)$, $\delta_2(x)$ are the roughness profiles of the lower and upper surfaces, respectively,

$\Lambda = 6\mu(u_1 + u_2)$ and μ is the viscosity of the lubricant,

$v = (u_2 - u_1)/(u_1 + u_2)$,

u_1 and u_2 are the rolling or sliding speeds of the lower and upper surfaces.

In the stochastic model, the roughness parameters $\delta_1 = \delta_1(x,\omega)$, and $\delta_2 = \delta_2(x,\omega)$ are random functions (or processes) of x for which the probability distributions are given. The goal is to compute the mean pressure $\langle p \rangle$ from (1) subject to the boundary conditions (2). It is a common practice to derive an equation for the mean pressure $\langle p \rangle$.

Consider the boundary-value problem (1) and (2). For each realization of δ_1 and δ_2, an integration of (1) yields

$$p(x) = M \int_0^x \frac{dy}{H^3(y)} + \Lambda \int_0^x \frac{h(y) + v[\delta_1(y) - \delta_2(y)]}{H^3(y)} dy \tag{3}$$

where by applying the boundary-condition (2) at $x = L$,

$$M = -\Lambda \int_0^L \frac{h(y) + v[\delta_1(y) - \delta_2(y)]}{H^3(y)} dy \Bigg/ \int_0^L \frac{dy}{H^3(y)} \; . \tag{4}$$

For convenience, let

$$\begin{aligned} \xi &= \delta_1 + \delta_2 \\ \eta &= \delta_1 - \delta_2 \end{aligned} \tag{5}$$

Then (4) becomes

$$M = -\Lambda \int_0^L \frac{h(y) + v\eta(y)}{H^3(y)} dy \Bigg/ \int_0^L \frac{dy}{H^3(y)} \; . \tag{6}$$

Suppose that the roughness profiles $\delta_1(x,\omega)$, $\delta_2(x,\omega)$, $x \geq o$, are random processes satisfying

(a) δ_1, δ_2 are continuous stochastic processes, defined for $x \geq 0$,

(b) the maximum $\max\limits_{o\leq x\leq L} \{|\delta_1(x)| + |\delta_2(x)|\}$ is less than the minimum h_m of the smooth film thickness, $h_m = \min\limits_{o\leq x\leq L} h(x)$, for almost every realization,

(c) the deviation $D(x) = |H^{-3}(x) - \langle H^{-3}(x)\rangle|$ satisfies the following condition of asymptotic weak dependence:

$$\lim_{L\to\infty} \frac{1}{L} \int_0^L \langle D(L)D(x)\rangle dx = 0 \; .$$

Since the random functions involved are bounded and positive. The Ergodic

Theorem stated on (p.177 [1]) becomes applicable to (6) under the above conditions (a) - (c). By rewriting the expression (6) and invoking the stated theorem, we have, with almost certainty

$$-M = \Lambda \frac{1}{L}\int_0^L \frac{h(y) + v\eta(y)}{H^3(y)}\, dy \Bigg/ \frac{1}{L}\int_0^L \frac{1}{H^3(y)}\, dy$$

$$\to \Lambda \int_0^L \left\langle \frac{h(y) + v\eta(y)}{H^3(y)} \right\rangle dy \Bigg/ \int_0^L \left\langle \frac{1}{H^3(y)} \right\rangle dy, \quad \text{as } L \to \infty \tag{7}$$

Noting (7) and averaging the equation (3), we differentiate the resulting equation twice to get

$$\frac{d}{dx}\left[\left\langle \frac{1}{H^3} \right\rangle^{-1} \frac{d\langle p \rangle}{dx} \right] = \Lambda \frac{d}{dx}\left[h + v \frac{\left\langle \frac{\delta_1}{H^3} \right\rangle - \left\langle \frac{\delta_2}{H^3} \right\rangle}{\left\langle \frac{1}{H^3} \right\rangle} \right], \tag{8}$$

which is our first announced result.

For a smooth bearing, the pressure p_o satisfies the equation

$$\frac{d}{dx}\left(h^3 \frac{dp_o}{dx} \right) = \Lambda \frac{dh}{dx}, \quad p_o(0) = p_o(L) = 0 \tag{9}$$

Subtracting (9) from (1), one obtains, noting (5)

$$\frac{d}{dx}\left(H^3 \frac{dp}{dx} - h^3 \frac{dp_o}{dx} \right) = \Lambda v \frac{d\eta}{dx} \ . \tag{10}$$

After adding and subtracting the term $H^3 \frac{dp_o}{dx}$ in the above parenthesis, the equation (10) can be written as

$$\frac{d}{dx}\left(H^3 \frac{d\delta p}{dx} \right) = \frac{dQ}{dx}, \quad \delta p(0) = \delta p(L) = 0 \tag{11}$$

where $\delta p = p - p_o$, is the deviation of p from the smooth case, and

$$Q = \Lambda v \eta - \left(H^3 - h^3 \right) \frac{dp_o}{dx} \ . \tag{12}$$

A simple integration of (11) yields

$$\delta p(x) = \int_0^x Q(y)H^{-3}(y)dy - \int_0^x H^{-3}(y)dy \, \frac{\int_0^L Q(y)H^{-3}(y)dy}{\int_0^L H^{-3}(y)dy} \; . \tag{13}$$

Splitting each of the integrals from 0 to L into two parts, from 0 to x and x to L, the equation (13) is reduced to

$$\delta p(x) = \frac{-\int_0^x H^{-3}dy \int_x^L QH^{-3}dy + \int_0^x QH^{-3}dy \int_x^L H^{-3}dy}{\int_0^L H^{-3}dy} \; . \tag{14}$$

It follows that

$$\begin{aligned} |\delta p(x)| &< \frac{\int_0^x H^{-3}dy \int_x^L |Q|H^{-3}dy + \int_x^L H^{-3}dy \int_0^x |Q|H^{-3}dy}{\int_0^L H^{-3}dy} \\ &\le \int_0^L |Q|H^{-3}dy \end{aligned} \tag{15}$$

By applying Schwarz's inequality, one gets the mean (absolute) deviation

$$\begin{aligned} \sigma_1 &= \max_{0\le x\le L} \langle |\delta p(x)| \rangle \\ &\le \left\{\int_0^L \langle Q^2(x)\rangle dx \int_0^L \langle H^{-6}(x)\rangle dx\right\}^{1/2} \end{aligned} \tag{16}$$

Similarly, squaring the expression (15), a bound for the mean-square deviation σ_2^2 is obtained

$$\begin{aligned} {\sigma_2}^2 &= \max_{0\le x\le L} \langle |\delta p|^2 \rangle \\ &\le \left\{\int_0^L \int_0^L \langle Q^2(x)Q^2(y)\rangle dxdy \int_0^L \int_0^L \langle H^{-3}(x)H^{-3}(y)\rangle dxdy\right\}^{1/2} \end{aligned} \tag{17}$$

The above bounds can be made explicit by estimating the correlation function $\langle Q^2 \rangle$ and $\langle Q^2(x)Q^2(y) \rangle$ in terms of the roughness correlations of ξ and η. For brevity, only the mean deviation σ_1 will be treated. Let $R_\xi = \langle \xi^2 \rangle$ and $R_\eta = \langle \eta^2 \rangle$. Then, after a sequence of algebraic inequalities, it can be shown that

$$\sigma_1 \leq \sqrt{2}A\left\{v^2 \int_0^L R_\eta(x)dx + 49 \int_0^L R_\xi(x) \left[\frac{\Delta h(x)}{h(x)}\right]^2 dx\right\}^{1/2} \tag{18}$$

$$\text{with } A = \Lambda\left\{\int_0^L \langle H^{-6}(x)\rangle dx\right\}^{1/2}$$

Let $|\delta W|$ denote the mean deviation of the load-carrying capacity defined as

$$|\delta W| = \int_0^L \langle|\delta p(x)|\rangle dx \tag{19}$$

Then, noting (16), the unit mean deviation ρ is given by

$$\rho = \left|\frac{\delta W}{W_o}\right| \leq \frac{\sigma_1 L}{\int_0^L p_o \, dx} \tag{20}$$

A similar bound for the unit mean-square deviation can be derived in terms of the fourth order correlations. For details and a concrete example, one is referred to [1].

III. <u>RELATED PROBLEMS</u>. The problem of a slider bearing of finite width is governed by the Reynolds equation in two dimensions. This partial differential equation was derived from the Stokes equations in hydrodynamics in three dimensions, based on the main assumption that the film thickness is small [2]. This is an example of many random boundary problems of "thin domain", that is, one component of the domain is small compared with others. In general, the boundary valued problem with an irregular domain is difficult to deal with. However, for a thin domain, the problem can often be reduced to one of regular domain, as exemplified by the Reynolds theory of hydrodynamic lubrication. To illustrate this principle, we consider, symbolically, the boundary-value problem

$$Lu = f \text{ in } D , \tag{21}$$

$$u|\partial D = g , \tag{22}$$

where L is an elliptic operator, such as a Laplacian, D is a random domain of two or higher dimensions with ∂D as its boundary, and f, g are given functions. Suppose that D is decomposible as $D = D_1 \times D_2$, such that the diameter $d(D_1)$ of the random set D_1 is much smaller than a characteristic length of the problem, but the component D_2 is regular. Then, for each realization, one can take a spatial average of (21) over the set D_1 and make use of the divergence, mean-value theorems or others in the integral calculus to incorporate the boundary condition on ∂D_1. This procedure often yields an approximate boundary-value problem in a reduced, regular domain D_2:

$$\tilde{L}\,(\partial D_1)\tilde{u} = \tilde{f} \text{ in } D_2 \quad , \tag{23}$$

$$\tilde{u}|\partial D_2 = \tilde{g} \ . \tag{24}$$

Here $\tilde{L}(\partial D_2)$ is an elliptic or ordinary differential operator with random coefficients resulting from the random part of the boundary ∂D_1.

For example, in the problem of the optical fiber as a wave guide, D_1 is the thin cross section of the fiber with random imperfections, and L designates the reduced wave operator. In this case, the system (23) and (24) constitutes two-point boundary-value problem for a random ordinary differential equation. As another example, the system (21) and (22) may describe the motion of long water waves over the ocean bed of a random topography. Since the wave length is much greater than the water depth, a reduced equation (23) similar to the Reynolds equation is two dimensions can be derived.

The reduced problem (23) - (24) is simpler than the original problem (21) - (22) in the sense that differential equations with random coefficients are better understood, in contrast with a random boundary problem. For instance there are more reliable methods of approximation in solving the random system (23) - (24) [3]. However, applications of such methods to specific problems will not be discussed here.

REFERENCES

[1]. P. L. Chow and E. A. Saibel, "On the Roughness Effect in Hydrodynamic Lubrication", ASME Journal of Lubrication Technology, Vol. 100, No. 1, 1978, pp. 176-180.

[2]. A. Cameron, Principles of Lubrication, Longmans, London (1966).

[3]. P. L. Chow, "Perturbation Methods in Stochastic Wave Propagation," SIAM Review, Vol. 17, No. 1, 1975, pp. 57-81.

A NEW MODEL FOR EVALUATING EFFECTIVENESS OF FRAGMENTING WARHEADS IN DYNAMIC ENCOUNTERS*

Edgar A. Cohen, Jr.
Applied Mathematics Branch
Naval Surface Weapons Center
White Oak Laboratory
Silver Spring, MD 20910

ABSTRACT. In this paper, we present a general methodology for assessing the probability of damage to a moving or stationary target due to a fragmenting warhead. The statistical theory developed is generally based on a "Poisson-Markov" model, with a Poisson process to describe warhead breakup and a Markov process to describe the effect of cumulative damage to the target. Our model relies on transformation of data procured through static testing into dynamic coordinates for a full three-dimensional encounter geometry. Shortcomings of this superposition procedure are also discussed.

I. INTRODUCTION. The purpose of this paper is to present a fundamental statistical methodology for assessing the damage to both moving and stationary targets due to a fragmenting warhead. At the time of burst of this warhead, it will generally be moving at some velocity with respect to an inertial frame. At issue is whether or not, in the situation wherein both the warhead and the target are moving, results from static testing can be used to predict, with reasonable accuracy, the effects in the dynamic engagement. We will be able to answer this question affirmatively provided that certain conditions are satisfied, to be spelled out clearly in the following paragraphs. The dynamic engagement differs from a static encounter in one essential detail, namely, the fact that the fragment and the target must obviously arrive at the same point at the same time. If the target remains stationary with respect to our inertial reference frame, we call such an encounter pseudodynamic. The target can then "wait for the fragment to arrive," so that the fragment need only travel in the proper direction to intercept its object. We will treat the process as a Poisson-Markov process, with a Poisson description for the target damage. The only philosophical objection would be to the assumption that the breakup is Poisson, which can only be strictly true under the supposition of independent increments [3, pg. 277]. Physically this means that fracturing occurs almost instantaneously, so that any relief waves set up in the material do not appreciably alter the fracture process. In any event, we shall assume an underlying Poisson process in our formulation of breakup.

II. USE OF STATIC ARENA RESULTS IN DYNAMIC PREDICTION. We proceed to build a dynamic engagement model based on information collected in static arena testing. As shown in Fig. 1 below, fragments are initiated by detonation of a warhead at the center of the arena, and they are typically collected in five degree polar zones. Various techniques are available for measuring velocities of

*This work was supported partly by the Systems Analysis Group, DG-10, NSWC and partly by the Independent Research Fund at NSWC.

the fragments [5, pp. 204-231], some of which may be more satisfactory than others. As far as their weight is concerned, the fragments are collected from bins, after which their masses may be measured directly.

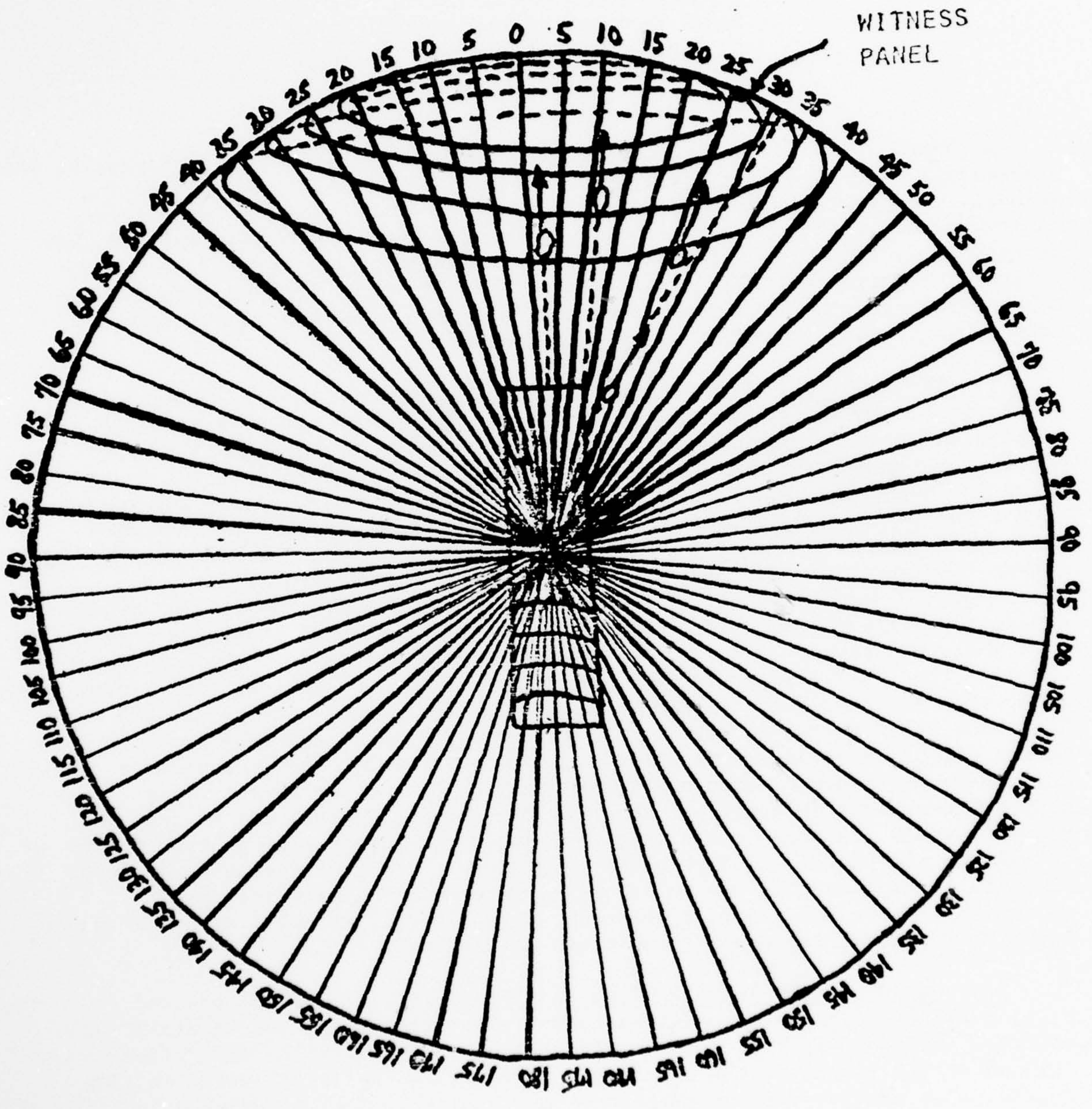


Fig. 1. Static Arena Test for Fragment Distribution

It is generally convenient to consider the warhead to be a distributed source of fragments, certainly for close encounters. For reasonably large distances with respect to the length of the bomb, one may want to consider it to be a point source. Of course, the model based on a distributed source is, of necessity, a bit more complicated, since one must be able to identify the initial location and flight direction of the fragment as it issues from the warhead surface.

In Fig. 2 we illustrate the general encounter geometry for the distributed source model. In this model, primed quantities represent dynamic warhead coordinates and unprimed quantities the static and target coordinates. For simplicity, we assume that the target point starts from a designated initial position and travels with a known constant velocity. The coordiuates shown in Fig. 2 are

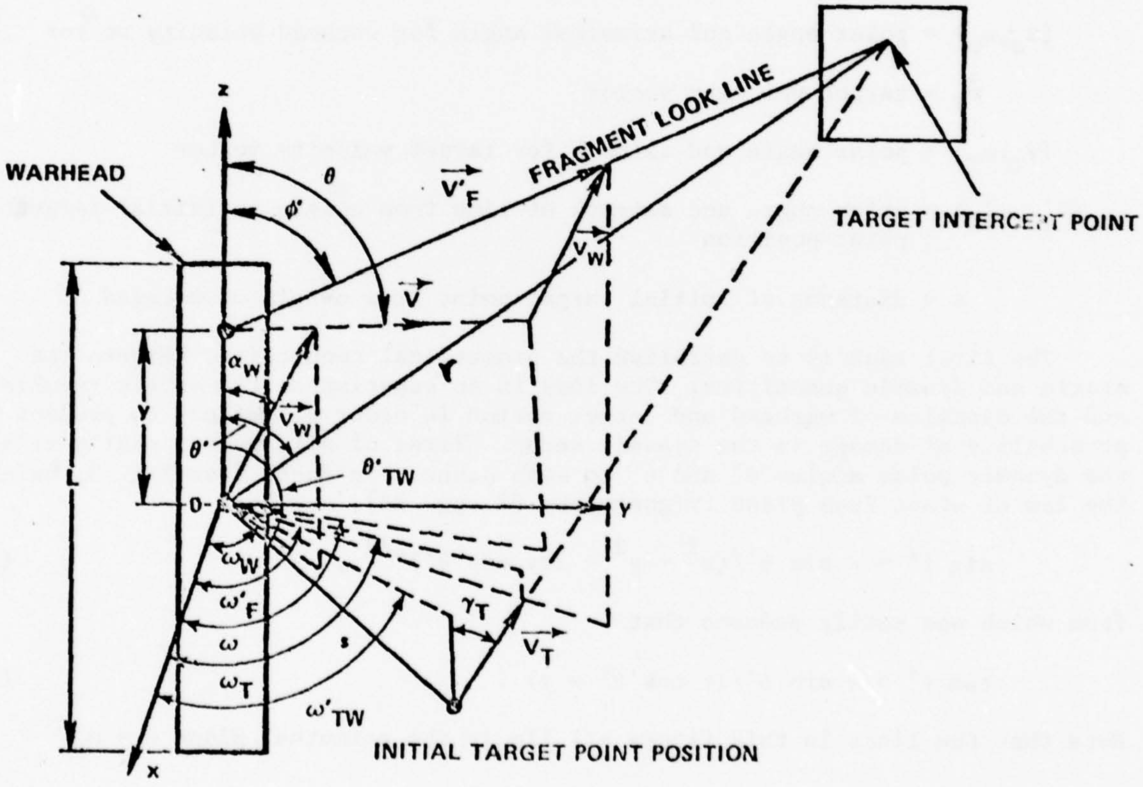


Fig. 2. Dynamic Encounter of Warhead and Target

(ϕ', ω_F') = dynamic polar angle and azimuthal angle of fragment look line to target intercept point

y = (possible) position of fragment relative to warhead origin

θ' = dynamic polar angle of target intercept point measured from warhead origin

r = distance at moment of burst from center of warhead to intercept point

$\vec{v}$ = static velocity vector

(ϕ, ω) = static polar angle and azimuthal angle of fragment look line

$\vec{v}_W$ = warhead velocity vector

(α_W, ω_W) = polar angle and azimuthal angle for warhead velocity vector

$\vec{v}_T$ = target velocity vector

(γ_T, ω_T) = polar angle and azimuth for target velocity vector

$(\theta_{TW}', \omega_{TW}')$ = polar angle and azimuth of line from origin to initial target point position

s = distance of initial target point from origin of warhead

The first task is to establish the geometrical connections between the static and dynamic quantities. The idea is to superimpose the static results and the dynamics of warhead and target motion in order ultimately to predict probability of damage in the dynamic setup. First of all, we can easily relate the dynamic polar angles θ' and ϕ' to each other. In fact, from Fig. 3, using the law of sines from plane trigonometry [2, pg. 88], one has

$$\sin \phi' = r \sin \theta' / (r^2 + y^2 - 2ry \cos \theta')^{1/2} , \qquad (1)$$

from which one easily deduces that

$$\tan \phi' = r \sin \theta' / (r \cos \theta' - y) . \qquad (2)$$

Note that the lines in this figure all lie in the azimuthal plane $\omega = \omega_F'$.

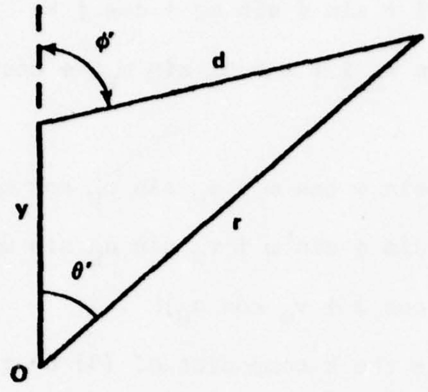


Fig. 3. Fragment Dynamic Polar Angle vs. Fiducial Dynamic Polar Angle

Now, to relate the static and dynamic quantities, let us appeal to Fig. 4, which shows a breakdown of static and dynamic warhead velocity-angular relationships.

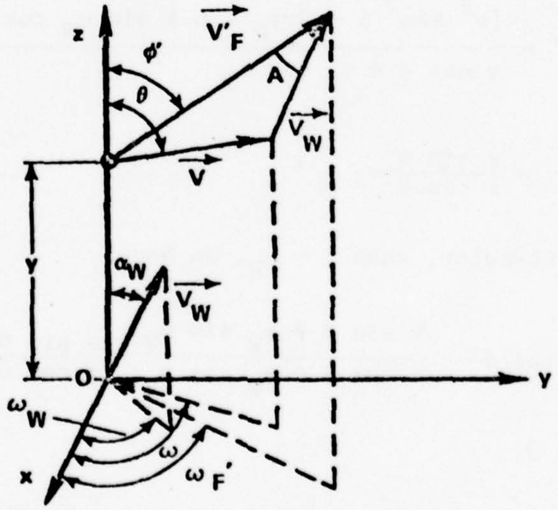


Fig. 4. Connection between Static and Dynamic Warhead Fragment Vectors

From this figure, one sees that

$$\vec{v} = v(\sin\phi\cos\omega i + \sin\phi\sin\omega j + \cos\phi\, k)$$

$$\vec{v}_W = v_W(\sin\alpha_W\cos\omega_W\, i + \sin\alpha_W\sin\omega_W j + \cos\alpha_W\, k),$$

so that

$$\begin{aligned}\vec{v}_F' = \vec{v} + \vec{v}_W = {} & (v\sin\phi\cos\omega + v_W\sin\alpha_W\cos\omega_W)i \\ & + (v\sin\phi\sin\omega + v_W\sin\alpha_W\sin\omega_W)j \\ & + (v\cos\phi + v_W\cos\alpha_W)k\,.\end{aligned} \tag{3}$$

To obtain $\cos\phi'$, one divides the k component of (3) by the magnitude of $\vec{v}_F'$ to obtain

$$\cos\phi' = \frac{v\cos\phi + v_W\cos\alpha_W}{\{v^2 + 2vv_W[\sin\phi\sin\alpha_W\cos(\omega - \omega_W) + \cos\phi\cos\alpha_W] + v_W^2\}^{1/2}}, \tag{4}$$

or

$$\begin{aligned}\tan\phi' &= \frac{[v^2\sin^2\phi + 2vv_W\sin\phi\sin\alpha_W\cos\ (\omega-\omega_W) + v_W^2\sin^2\alpha_W]^{1/2}}{v\cos\phi + v_W\cos\alpha_W} \\ &= \frac{r\sin\theta'}{r\cos\theta' - y}\,,\end{aligned} \tag{5}$$

by (2). In particular, when $\omega = \omega_W$, we have

$$\tan\phi' = \frac{v\sin\phi + v_W\sin\alpha_W}{v\cos\phi + v_W\cos\alpha_W} = \frac{r\sin\theta'}{r\cos\theta' - y}\,, \tag{6}$$

and, when $\alpha_W = 0$,

$$\tan\phi' = \frac{v\sin\phi}{v\cos\phi + v_W} = \frac{r\sin\theta'}{r\cos\theta' - y}\,. \tag{7}$$

Let us suppose now that the radius of the static arena is R, and let θ be the static polar angle measured from the warhead center, as illustrated in Fig. 5.

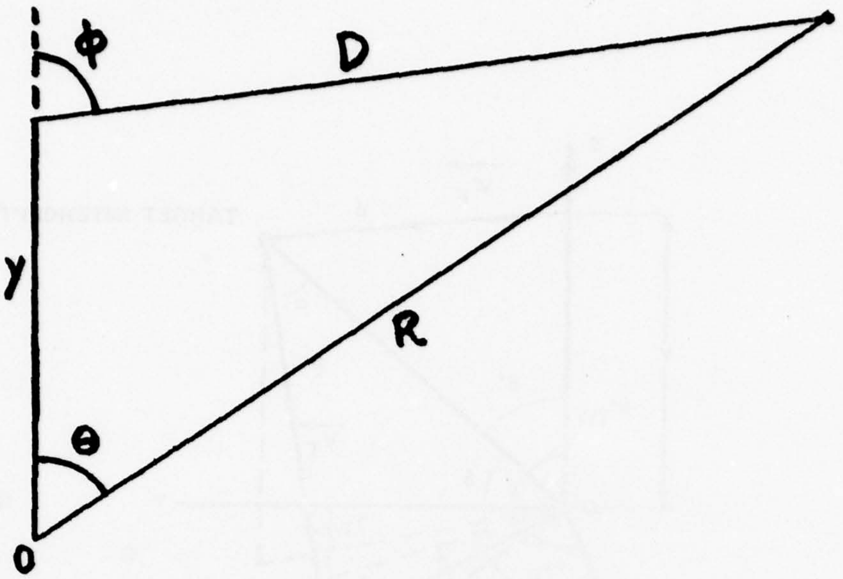


Fig. 5. Relationship between Fragment Static Polar Angle and Static Arena Polar Angle

Then, by the law of sines and the law of cosines for plane triangles [2, pp. 88-90], it follows that

$$\frac{\sin \phi}{R} = \frac{\sin \theta}{D} , \tag{8}$$

where

$$D^2 = y^2 + R^2 - 2yR \cos \theta .$$

Equations (4) and (5), together with (8), give an important connection between results in the dynamic arena and the static arena in order that a given target point be intercepted. From Fig. 6, one can also obtain an expression for r, the distance between the center of the warhead and the target intercept point. First of all, by taking the dot product of the unit vector in the r direction with that in the $\vec{v}_T$ direction, one obtains cos α. Likewise, if a unit vector in the s direction (the direction from the origin to the target point position at burst) is dotted with that in the $\vec{v}_T$ direction, - cos β is obtained. Therefore, by the law of sines,

$$r = s \sin \beta / \sin \alpha , \tag{9}$$

where

$$\cos \alpha = \sin \theta' \cos \omega_F' \sin \gamma_T \cos \omega_T + \sin \theta' \sin \omega_F' \sin \gamma_T \sin \omega_T + \cos \theta' \cos \gamma_T$$

and

$$- \cos \beta = \sin \gamma_T \cos \omega_T \sin \theta_{TW}' \cos \omega_{TW}' + \sin \gamma_T \sin \omega_T \sin \theta_{TW}' \sin \omega_{TW}'$$
$$+ \cos \gamma_T \cos \theta_{TW}' .$$

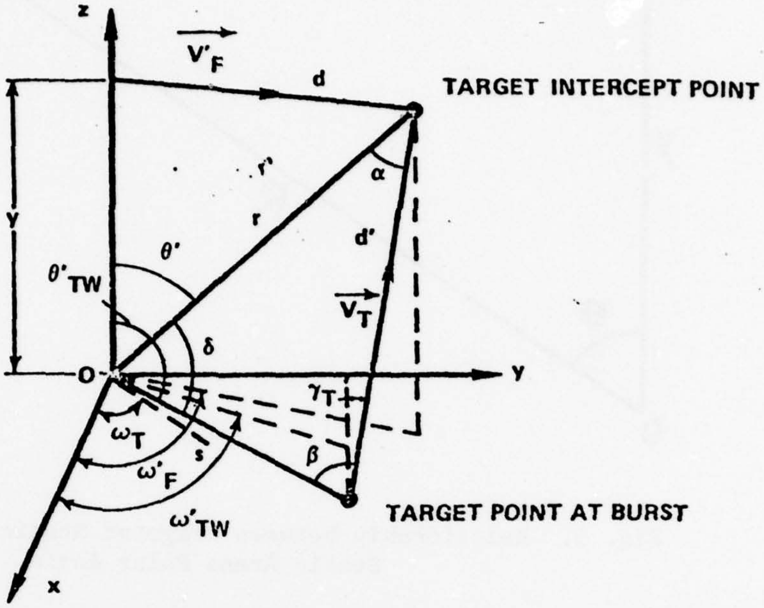


Fig. 6. Fragment and Target Encounter Geometry

III. USE OF DRAG EQUATION TO PREDICT TERMINAL VELOCITY. In section II, we derived geometrical relations which must be satisfied in order that target interception occur in the dynamic case. We mentioned in section I that, in the dynamic mode, one must also be certain that a fragment starting from a given position y and traveling in a proper direction will actually intercept the target point. In other words, there is a certain time requirement present in a dynamical situation which is not present in either a static or a pseudodynamic encounter. It is necessary now to study in some detail the dynamics of the motion of both the fragment and its intended target. First of all, let us consider the classical drag equation, namely [5, pg. 210],

$$D = \frac{1}{2} C_D \rho A_p v^2 = - M_0 dv/dt, \qquad (10)$$

where

D = drag force

C_D = drag coefficient

ρ = air density

A_p = area presented to the flow

v = speed at time t.

In using (10), we neglect the force of gravity, which we consider to be a second order effect compared with aerodynamic forces. (10) can then be used to predict the speed, since any forces orthogonal to the direction of motion do not affect the magnitude of the velocity vector. In the simplest case, one may assume, for short distances, straight line motion. Also, generally speaking, the drag coefficient C_D depends on presented area, Mach number, and Reynolds number [1, pp. 243-244], but, if the fragment flies supersonically (e.g., in the range of Mach 3 to Mach 6), we may assume, for simplicity, that C_D depends only on presented area A_p. In other words, the fragment is in an unguided flight regime, where change in drag coefficient occurs because of the tumbling motion. Now, instead of time, if we consider distance along the path, (10) can be replaced by an equivalent linear differential equation, namely,

$$dv/dr_1 = \frac{1}{2} C_D \rho A_p v/M_0 \ . \tag{11}$$

Equation (11) can be explicitly solved for remaining speed v_r to give

$$v_{r_1} = v_F' \ \exp\left[- \rho \int_0^{r_1} C_D(A_p(s))A_p(s)ds/2M_0\right] \ , \tag{12}$$

where

r_1 = distance traveled by fragment

v_F' = initial dynamic speed of fragment.

Therefore, since $dt/dr_1 = 1/v_{r_1}$, the time of flight of the fragment is given by

$$T_F = \frac{1}{v_F'} \int_0^d \exp\left[\rho \int_0^{r_1} C_D(A_p(s))A_p(s)ds/2M_0\right]dr_1 \ , \tag{13}$$

where d is the distance the fragment must travel, as indicated in Fig. 6. For target interception to occur, T_F must be equated to the time of flight of the target, namely, d'/v_T. The quantities d and d' can be readily obtained via the target geometry. One sees from Fig. 6, in fact, that

$$d = (y^2 + r^2 - 2ry \cos \theta')^{1/2}$$

and that

$$d' = (r^2 + s^2 - 2rs \cos \delta)^{1/2} \ .$$

The quantity $\cos \delta$ is obtainable as the dot product of a unit vector in the r direction with that in the s direction, namely,

$$\cos \delta = \sin \theta' \cos \omega_F' \sin \theta_{TW}' \cos \omega_{TW}' + \sin \theta' \sin \omega_F' \sin \theta_{TW}' \sin \omega_{TW}' + \cos \theta' \cos \theta_{TW}' \ . \tag{14}$$

It follows that T_F is well-defined by the target geometry alone, and it remains to be seen how (13) can be satisfied. To see this, we must first assume a form for the drag coefficient as a function of the presented area. The simplest form is a linear function of the area, i.e., a first order Taylor series expansion. We must also assume a known rate of change of presented area with distance s and a known initial presented area $A_p(0)$. In other words, we assume that

$$\begin{cases} dA_p(s)/ds = g(s), \\ A_p(0) \text{ given.} \end{cases} \tag{15}$$

The function g(s) is some known function of s. The question naturally arises as to how g(s) is to be determined. If we can suppose that $g(s) \equiv g(0)$, i.e., that the initial tumbling rate is, for all practical purposes, preserved, then we are saying, in effect, that the translational rate and the tumbling rate can be decoupled, with one having little or no effect on the other. The extent to which this is or is not true affects whether or not we can get by with static testing alone. In other words, are the initial conditions of the problem adequate to determine the behavior, or is there a complex interrelationship between fragment velocity and fragment angular momentum to be taken into account?

Going back to Equation (13) and supposing that

$$C_D(A_p(s)) = a_1 + a_2\ A_p(s), \tag{16}$$

where a_1 and a_2 are empirically determined quantities depending on the shape of the fragment, we find that

$$T_F = \frac{1}{v'_F} \int_0^d \exp[\rho \int_0^{r_1} [a_1 + a_2\ A_p(s)]A_p(s)ds/2M_0]dr_1, \tag{17}$$

where, from (15),

$$A_p(s) = \int_0^s g(x)dx + A_p(0). \tag{18}$$

Generally speaking, (17) becomes a functional of g(x) and the initial fragment attitude $A_p(0)$, i.e., a quantity of the form

$$T_F = G[g(x),\ A_p(0)]. \tag{19}$$

In case $g(x) \equiv g(0)$, (19) reduces to a two-parameter family of initial angular rates and angular orientations. Once such a family of pairs is obtained, we can determine the residual speed corresponding to any such couple by using (12), with r replaced by d.

To summarize up to this point, in a general dynamic encounter, certain static-dynamic vector relationships exist. The density function for a hit at some dynamic polar angle θ' on a given target point P is obtainable, at least in theory, as a multidimensional integral of the density over the appropriate set. We denote this density by $f_{\theta'}$. (5) shows, for example, that, in general, azimuth ω is involved in determining the relation between ϕ' and the variables in the static arena. However, in static testing, one does not normally distinguish among fragments with different azimuths. Fragments are grouped only according to polar zones. The general case can only be handled if one broadens the concept of a static test to include azimuthal as well as polar zones. However, in the case where $\vec{v}_W$ aligns with the warhead axis (which may very well be the usual situation in practice), this is not necessary, and we can also assert, from considerations of symmetry, that, for a given polar angle θ' and azimuth angle ω, the probability density for a hit is just $f_{\theta'}/2\pi$. In other words, we need not distinguish among fragments which follow different azimuths for the same polar angle.

IV. THE POINT SOURCE MODEL OF A DYNAMIC ENCOUNTER. In our static tests, it may be inconvenient to assume that the warhead is a distributed source of fragments, since this means that testing must also include initial location identification of fragments with respect to the chosen warhead origin as well as velocity-polar angle relationships. If conditions are such that we can treat the warhead as a point source (e.g., if the distances involved are great enough relative to the size of the warhead), then we obviate the necessity of locating the initial positions of the fragments, and, at the same time, we no longer need worry about the angular relationship between the warhead velocity vector and the warhead axis. From a practical point of view, one can see the advantages of treating the missile as a point source when one can.

For this simpler model, let us refer to Fig. 7. In the context of the point source, one may align the Z-axis with the $\vec{v}_W$ velocity vector. Either by reference to Fig. 7 or by substitution of y = 0 into (7), we find that

$$\tan \theta' = v \sin \phi/(v \cos \phi + v_W). \qquad (20)$$

Again we must obtain the condition that the fragment intercept the target. As one sees from Fig. 8, the time required for the target to traverse distance d' is just d'/v_T, where

$$d' = s \sin B/\sin A$$

and

$$\begin{aligned} \cos A &= \sin \theta' \cos \omega \sin \gamma_T \cos \omega_T + \sin \theta' \sin \omega \sin \gamma_T \sin \omega_T \\ &\quad + \cos \theta' \cos \gamma_T \qquad (21) \\ \cos B &= \sin \theta' \cos \omega \sin \theta'_{TW} \cos \omega'_{TW} + \sin \theta' \sin \omega \sin \theta'_{TW} \sin \omega'_{TW} \\ &\quad + \cos \theta' \cos \theta'_{TW} . \end{aligned}$$

Equations (21) are obtained by considering dot products of appropriate unit vectors.

Relation (13) again gives the time required for the fragment to traverse d, where, referring to Fig. 8,

$$d = d' \sin C/\sin B$$

and

$$-\cos C = \sin \theta'_{TW} \cos \omega'_{TW} \sin \gamma_T \cos \omega_T + \sin \theta'_{TW} \sin \omega'_{TW} \sin \gamma_T \sin \omega_T$$
$$+ \cos \theta'_{TW} \cos \gamma_T. \tag{22}$$

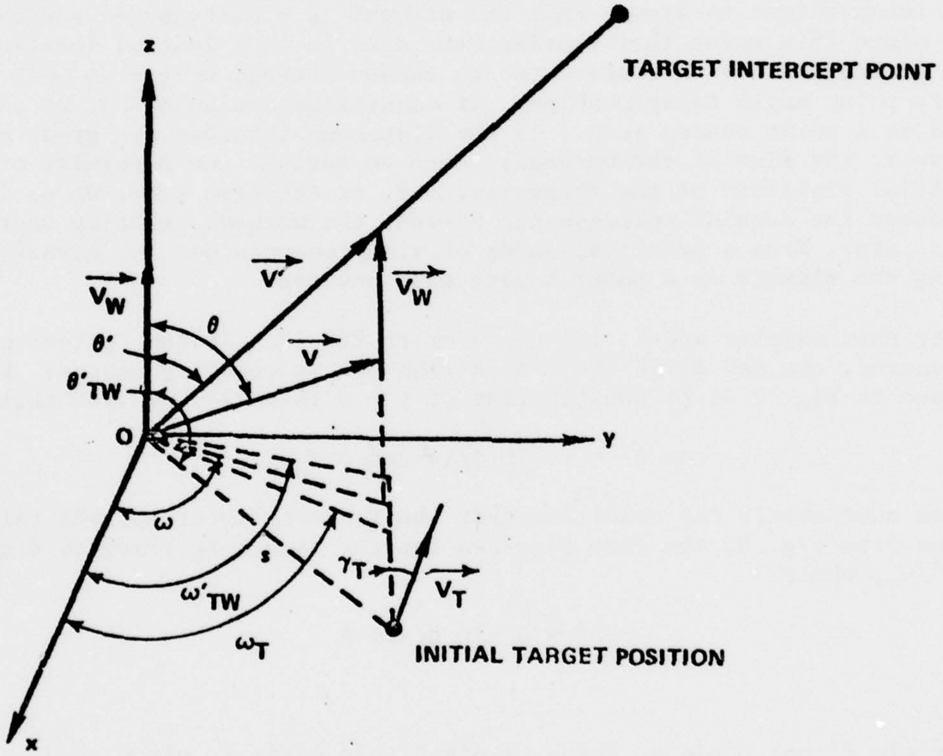


Fig. 7. Point Source Model of Dynamic Encounter of Warhead and Target

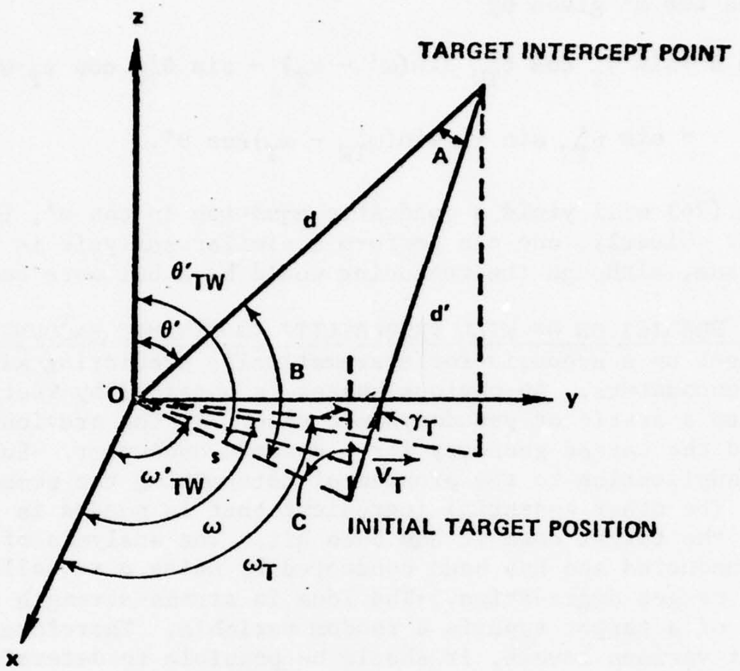


Fig. 8. Fragment and Target Encounter Geometry for Point Source

In addition, from Fig. 7, one sees that

$$v^2 = v_F'^2 + v_W^2 - 2v_F' v_W \cos \theta' \tag{23}$$

and that

$$\sin(\theta - \theta') = v_W \sin \theta'/v. \tag{24}$$

Of course, the azimuth of $\vec{v}$ is the same as that for $\vec{v}_F'$ in the point source model. Another point can be made here and that is the following: If the dynamic polar angle θ' is specified, together with the initial target location and target velocity vector, one can generally expect a dependence of the azimuth ω' on the other parameters in order that target interception occur. First of all, consider the cross product of a unit vector in the initial target direction and a unit vector in the $\vec{v}_T$ direction. This gives

$$\begin{aligned} \vec{W} &= (\sin \theta_{TW}' \sin \omega_{TW}' \cos \gamma_T - \sin \gamma_T \sin \omega_T \cos \theta_{TW}')i \\ &+ (\sin \gamma_T \cos \omega_T \cos \theta_{TW}' - \sin \theta_{TW}' \cos \omega_{TW}' \cos \gamma_T)j \\ &+ \sin \theta_{TW}' \sin \gamma_T \sin (\omega_T - \omega_{TW}')k, \end{aligned} \tag{25}$$

a vector normal to the plane of $\vec{s}$ and $\vec{v}_T$. Such a vector must be orthogonal to the $\vec{v}_F'$ vector, since $\vec{v}_F'$ lies in the plane of $\vec{s}$ and $\vec{v}_T$. This leads to the condition for ω' given by

$$\sin\theta'[\sin\gamma_T \cos\theta_{TW}' \sin(\omega' - \omega_T) - \sin\theta_{TW}' \cos\gamma_T \sin(\omega' - \omega_{TW}')]$$

$$= \sin\theta_{TW}' \sin\gamma_T \sin(\omega_{TW}' - \omega_T)\cos\theta' . \qquad (26)$$

Relation (26) will yield a quadratic equation in $\cos\omega'$, from which ω' can be obtained. Clearly, one can perform a similar analysis in the distributed source case, although the reasoning would be a bit more complicated.

V. PREDICTION OF KILL PROBABILITY IN DYNAMIC ENCOUNTERS. In this section we will set up a scenario for systematically predicting kill probabilities in dynamic encounters. An obvious subset is obtained by restricting oneself, of course, to a static or pseudodynamic case. In the previous section, we have presented the target geometry for a dynamic encounter. Such geometry has obvious application to the problem of determining the probability of hitting the target. The other essential ingredient that is needed is the probability of damaging the target once it has been hit. The analysis of the latter problem can be conducted and has been conducted by using a so-called stress-strength model of target degradation. The idea in stress-strength modeling is that the strength of a target type is a random variable. Therefore, by stressing the target at various levels, it should be possible to determine the distribution of strengths. It is necessary then to construct either an empirical relationship or possibly an analytic relationship which adequately reflects this stress as a function of the inputs. Various formulas have been proposed, one of which is based on the ratio of impact speed to a quantity called "critical velocity." The term critical velocity is interpreted by vulnerability analysts in several ways, some calling it that velocity above which half of the time the target is defeated. Others choose to define it as that velocity above which, in a small sample of tests, all targets are destroyed. One such empirical formula for critical velocity is given by

$$v_c = k_1 e^{k_2\alpha + k_3(t/k_4 w^{1/3})}/k_4^{1/2} w^{1/6} , \qquad (27)$$

where k_1, k_2, k_3, and k_4 are shape-dependent parameters to be determined and

t = target shield thickness

w = fragment weight

α = obliquity angle (the angle between the impact velocity vector and the normal to the target at the impact point).

The stress Ω is then just the ratio

$$\Omega = v_I/v_c . \qquad (28)$$

One then associates with Ω another quantity F, the cumulative probability function for damage, so that $F = g(\Omega)$. If one is able to build a good functional correspondence between damage level and probability of damage, then this should be reflected in the fact that there will be a 1-1 correspondence between level surfaces of Ω and probability of damage F. In this analysis, one may employ, for example, the Weibull distribution [4, pp. 184-258], together with the maximum likelihood method [4, pg. 81].

Now the time of flight t_F of a fragment which impinges on a dynamic target point P from a given position y is, by the discussion in section III, identified with a state vector $\vec{u}_{P,y} = (\theta', \omega_F', M, v_F', \dot{A}_P, A_P(0), S; P,y)$. Keeping this fact in mind, we propose the following scenario for kill probability prediction:

1. Let $f_1(t_F(\vec{u}_{P,y}))$ represent the probability density of target interception on dynamic point P for the given state $\vec{u}_{P,y}$.

2. Using results of Section III, the dynamic impact velocity $\vec{v}_I'$ of the fragment can be determined. Let $\vec{v}_R' = \vec{v}_I' - \vec{v}_T$ represent the residual velocity between the velocity of impact and the target velocity. The obliquity angle α alluded to previously in this section is then the angle between $\vec{v}_R'$ and the target normal at P.

3. Use an appropriate damage criterion to obtain the cumulative damage probability, given a hit on the assigned target point P while in state $\vec{u}_{P,y}$. We shall denote this probability by $F(D|\vec{u}_{P,y})$.

4. The probability density $f_2(y,P)$ for damage from y on target point P can be compactly written as

$$f_2(y,P) = \int_U F(D|\vec{u}_{P,y}) f_1(t_F(\vec{u}_{P,y})) d\vec{u}_{P,y} ,$$

where U is the set of possible states $\vec{u}_{P,y}$.

5. Integrate $f_2(y,P)$ over the warhead and target to obtain the probability $P(N)$ of target damage for a given number N of fragments:

$$P(N) = \iint f_2(y,P) dydP .$$

6. Letting $Q(N)$ = probability of getting N fragments, compute the unconditional kill probability:

$$P_K = \sum_{N=1}^{\infty} Q(N) P_K(N) . \tag{29}$$

If the quantity Q(N) is to be computed on the basis of a Poisson breakup process, we would suppose that [3, pg. 140]

$$Q(N) = e^{-kA_s} \frac{(kA_s)^{N-1}}{(N-1)!} , \qquad (30)$$

where A_s is the surface area of the warhead, so that k A_s+1 is the expected number of fragments. The problem arises as to how to compute $P_K(N)$, the kill probability based on N fragments. If the target could be instantaneously repaired, this probability can be thought of as the probability of at least one lethal hit, based on an independence of effects model. We then have

$$P_K(N) = 1-(1-P(N))^N . \qquad (31)$$

Surely Equation (31) gives a conservative estimate of target kill, since, for a model based on cumulative damage, the kill probability could only be enhanced. In other words, (29), together with (31), gives a conservative predictor for P_K. In the next section, we take into account more carefully the effect of cumulative damage. It is to be noted that (29) is not, generally speaking, a closed-form expression.

VI. THE MARKOV PROCESS FOR TARGET DAMAGE. A more sophisticated assessment of target damage can be obtained by assuming a Markov process for degradation. For this purpose, assume that the target is made up of r components, corresponding to 2^r possible damage states (including the undamaged state, of course). Damage states correspond to incapacitation of any number of the r components. Certainly there exists a subset C of these damage states which is critical, i.e., if the system is in any of the critical damage states, it is considered to be killed. Also, in a dynamic engagement, a Markov process should be inhomogeneous [3, pg. 251], since the distance to the target impact point changes with time, and this influences the velocity of impact relative to the target.

Let us order the possible arrival times of the N fragments, namely, $t_{(1)} \leq t_{(2)} \leq t_{(3)} \leq \cdots \leq t_{(N)}$. Now a fragment initially in state $\vec{u}_{p,y}$ hits the target point p with probability density $f_1(t_F(\vec{u}_{p,y}))$, as we have previously noted. For any other point z on the warhead, one can likewise determine the density $f_1(t_1(\vec{u}_{p,z}))$, corresponding to another state $\vec{u}_{p,z}$. By integration of f_1 over such states, where $t_1 \leq t_F$, we obtain the cumulative probability $F(t_F)$. Then the probability density for the $t_{(k)}$ order statistic is just [3, pg. 376]

$$f_{t_{(k)}}(t_F(\vec{u}_{p,y})) = \frac{N!}{(k-1)!(N-k)!} [F(t)]^{k-1}[1-F(t)]^{N-k} \cdot f_1(t_F(\vec{u}_{p,y})) . \qquad (32)$$

One proceeds again through steps 1 to 6, after which the Markov matrices [3, pg. 252], for $1 \leq k \leq N$,

$$M_k = (P_{ij}^{(k)})_{1 \leq i \leq 2^r,\ 1 \leq j \leq 2^r},\ P_{ij}^{(k)} = 0,\ i > j\ , \tag{33}$$

are formed. The product $\prod_{k=1}^{N} M_k$ gives the transition matrix for N fragments, which we denote by

$$T = (P_{ij}(N)),\ P_{ij}(N) = 0,\ i > j\ . \tag{34}$$

Since C represents the set of critical damage states, one has finally

$$P_k(N) = \sum_{j \epsilon C} P_{1j}(N), \tag{35}$$

where it is assumed that state 1 represents the undamaged state. $P_k(N)$, as given by (35), is then inserted into (29) in place of relation (31). We have therefore determined the unconditional kill probability.

SUMMARY. In this paper, we have presented a systematic theory for predicting kill probabilities in dynamic encounters of warheads with their targets. Although it is clear that the emphasis was on this dynamic interaction, obviously the theory is valid a fortiori for static or pseudodynamic engagements, in which either the warhead or the target or both are stationary. We have presented a model which could be construed as that of compounding a Markov process for predicting damage with that of a Poisson process for predicting fragment formation. Also, we have given a first order model, based on the notion of independence of effects, which should afford a lower bound for the actual kill probability, since association of effects is then ignored. We have not yet included in this model so-called synergistic effects, wherein there is constructive reinforcement of two or more fragments impinging on a target at almost the same time. This is a topic for further research. It is hoped that the model can serve as a useful guide for improvement of the prediction of kill probabilities even though it may not be numerically implementable to the last detail.

ACKNOWLEDGMENTS. The author would like to thank a number of people for their encouragement in this work, in particular, Messrs. Jim Ball, Tom McCants, Jim Logan, Fred Clodius, and Joe Colick, of the Systems Analysis Group at NSWC. He would also like to thank Messrs. William McDonald and Ted Orlow, with whom he has had frequent fruitful discussions.

REFERENCES

1. Arthur, Wallace and Fenster, Paul K., Mechanics, Holt, Rinehart and Winston, Inc., 1969

2. Ayres, Frank, Jr., Schaum's Outline of Theory and Problems of Plane and Spherical Trigonometry, Schaum Publishing Co., 1954

3. Fisz, Marek, Probability Theory and Mathematical Statistics, John Wiley & Sons, Inc., New York, London

4. Mann, Nancy R., Schafer, Ray E., and Singpurwalla, Nozer D., Methods for Statistical Analysis of Reliability and Life Data, John Wiley and Sons, Inc., 1974

5. McShane, Edward J., Kelley, John L., and Reno, Franklin V., Exterior Ballistics, University of Denver Press, 1953.

GENERATING THE EFFICIENT SET FOR MULTIPLE OBJECTIVE LINEAR PROGRAMS*

J. G. Ecker
Mathematical Sciences Department
Rensselaer Polytechnic Institute
Troy, N.Y. 12181

Nancy S. Hegner
School of Business
State University of New York
Albany, N.Y. 12222

ABSTRACT. A method for generating all efficient points for linear multiple objective programs is presented. In the method we first provide a systematic procedure for determining whether or not any efficient points exist. If the set of efficient points is not empty then an initial efficient extreme point is determined. Given an efficient vertex, the method then generates all efficient vertices as well as all maximal efficient faces. The efficient set is then described as the union of faces.

1. INTRODUCTION. In a multiple objective linear program a convex polyhedron X is given over which several linear objectives are to be maximized. These objectives can be given as the components of a column Cx where C is a k x n matrix with k denoting the number of objectives. A point $x^o \in X$ is called efficient if there is no $x \in X$ with $Cx \geqq Cx^o$ and $Cx \neq Cx^o$. Efficient points are often called Pareto optimal points or nondominated points.

In this paper, we summarize a systematic procedure for generating the entire set that has been developed in references [2], [3], [4], and [5]. For other approaches to this problem see the selected references [11], [6], [13], [7], and [10]. References [1], [12], and [14] are good sources for a rather broad range of applications.

Throughout this paper we will consider a multiple objective problem of the form

P: max Cx subject to $x \in X$

and we will let E denote the set of efficient points.

2. FINDING AN INITIAL EFFICIENT EXTREME POINT Given a feasible point $x^o \in X$, to determine whether or not x^o is efficient

* Complete details of this presentation can be found in references [2], [3], [4], and [5].

it is natural to consider the linear program

$$\begin{aligned} Px^o: \quad & \underline{\max}\ e^T s \\ & \underline{\text{subject to}}\ Cx = Is + Cx^o \\ & x \varepsilon X,\ s \geqq 0 \end{aligned}$$

where $e^T = (1,1,\ldots,1)$. The following lemma is well known, see [6] for example.

Lemma 1. If $x^o \varepsilon X$ then $x^o \varepsilon E$ if and only if Px^o has a maximum value of zero.

Another characterization of efficiency is given by considering the so called weighted linear program,

$$P\lambda : \underline{\max}\ \lambda^T Cx\ \underline{\text{subject to}}\ x \varepsilon X.$$

Lemma 2. If $x^o \varepsilon X$ then $x^o \varepsilon E$ if and only if there is a $\lambda > 0$ such that x^o is optimal for $P\lambda$.

Proof. For a nice proof of this result, see [9].

Actually solving program Px^o gives much more information about E than originally suspected. The following two theorems due to Ecker and Kouada, [2], show that if $E \neq \emptyset$ then by solving Px^o an efficient point can always be found.

Theorem 1. Given $x^o \varepsilon X$, if $(\bar{x},\bar{s})$ is optimal for Px^o then $\bar{x} \varepsilon E$.

Proof. see [2].

We should remark that if $(\bar{x},\bar{s})$ solves Px^o then $\bar{x}$ need not be an extreme point of X. However, if $\bar{x}$ is interior to a face of X then that entire face is efficient and in [4], Ecker and Hegner show how to pivot on the optimal tableau for Px^o to obtain an efficient extreme point.

Theorem 2. Given $x^o \varepsilon X$, if Px^o has no finite maximum then $E = \emptyset$.

Proof. Suppose $\bar{x} \varepsilon E$. By Lemma 2, there is a vector $\lambda \geqq e$ such that $\lambda^T C\bar{x} \geqq \lambda^T Cx$ for each $x \varepsilon X$. If (x,s) is any feasible point for Px^o, we then have

$$e^T s \leqq \lambda^T s = \lambda^T (Cx - Cx^o) = \lambda^T Cx - \lambda^T Cx^o \leqq \lambda^T C\bar{x} - \lambda^T Cx^o$$

But if $e^T s$ is bounded above by $\lambda^T C\bar{x} - \lambda^T Cx^o$ then Px^o must have a finite maximum. This contradiction implies that $E = \emptyset$. (For an alternate proof, see [2]).

The above results show that simply by solving Px^o we can either obtain an efficient point or show that there are none.

3. GENERATING EFFICIENT FACES. Suppose we solve Px^o and obtain an initial efficient extreme point. Let the tableau T below represent the efficient extreme point x_T,

$$T \quad \begin{array}{c|c|cc|} & & x^B & x^N \\ \hline & d & -C & 0 \\ \hline & b & A & I \\ \hline \end{array} .$$

Here x^B denotes the variables that are basic at the initial vertex and x^N denotes the nonbasic variables. Notice that we have eliminated the nonbasic variables from the objectives and that we have recorded the negative of the objectives. (Technically we should distinguish the C in Tableau T from the initial matrix C defining the objectives. The initial matrix C has been transformed by new operations so that only coefficients of basic variables appear. Henceforth, C will refer to the matrix C in T.) From tableau T, the initial efficient extreme point x_T is given by

$$x_T = (x^B, x^N) = (b, 0).$$

The efficient set is connected (see [13]) and, in fact, given any two efficient vertices there is a path of efficient edges of X that connect the two vertices. In this section, we give an important characterization for an edge incident to x_T to be efficient. At the same time (and more generally), we give a characterization for faces incident to x_T to be efficient.

We will assume throughout the remainder of this presentation that X is bounded and nondegenerate. See [8], for a discussion of how these assumptions can be relaxed.

Given T representing the initial extreme point $x_T \in E$, let N_T denote the set of nonbasic indices. Each subset $F \subseteq N_T$ determines a face of X incident to x_T. Let

$$f(T,F) = \{x \in X | x_j = 0 \text{ for } j \in N_T - F\}$$

be the face incident to x_T obtained by letting all nonbasic variables be zero excepting those in F. For example, $f(T,\{j\})$ for $j \in N_T$ denotes the edge incident to x_T obtained by increasing the nonbasic variable x_j from zero and adjusting the basic variables in T to maintain feasibility. Note that if $F \subseteq F^*$ then $f(T,F) \subseteq f(T,F^*)$. For $F \subseteq N_T$, let

$$G(F) = \{(v,w) \geqq 0 | C^T v + w = -C^T e, \ w_j = 0 \text{ for } j \in F\} ,$$

where the components of w are indexed by the elements of N_T. Notice that G(F) depends only on the matrix C in T. The following is a major result in our method for generating E.

Theorem 3. The face f(T,F) is efficient if and only if $G(F) \neq \emptyset$.

Proof. see [5]. (The proof hinges on the fact that if $(\bar{v},\bar{w}) \varepsilon G(F)$ then f(T,F) is the optimal set for $P\lambda$ where $\lambda = \bar{v} + e$. See theorem 4 below.)

As a corollary to this theorem, we note that the edge incident to x_T obtained by increasing the nonbasic variable x_j is efficient if and only if there is a nonnegative solution (v,w) to

$$S_T : \; C^T v + w = -C^T e$$

with $w_j = 0$.

In view of theorem 3, we wish to find subsets $F \subseteq N_T$ with $G(F) \neq \emptyset$ and such that no $F^* \subseteq N_T$ exists with $G(F^*) \neq \emptyset$ and $F \subseteq F^*$ properly. The variable w_j is called nonredundant in G(F) if $G(F \cup \{j\}) \neq \emptyset$.

It is important to note that we need only consider faces incident to x_T having edges incident to x_T that are efficient since an efficient face must have all of its edges efficient. Thus we let

$$J_T \equiv \{j \;\varepsilon\; N_T | G(\{j\}) \neq \emptyset\}$$

and so J_T denotes the indices in N_T corresponding to efficient edges incident to x_T.

Subroutine FACE as described precisely in [5], uses the set J_T to systematically construct maximal subsets $F \subseteq J_T$ with $G(F) \neq \emptyset$ and thus by Theorem 3 we know that $f(T,F) \subset E$. Subroutine EDGE as described precisely in [3] starts by computing J_T and then systematically generates all efficient vertices by exploiting the fact that any two efficient vertices are connected by a path of efficient edges. In [5], subroutine EDGE and FACE are combined to form an algorithm for generating all efficient vertices and all maximal efficient faces. A crucial key in the algorithm is the ability to recognize previously generated maximal efficient faces. In the algorithm each maximal efficient face has a double describtion; it is described as a face f(T,F) incident to some efficient vertex x_T and it is also described by identifying a vector λ so that f(T,F) is the optimal set for the weighted linear program

$$P\lambda : \underline{\max}\ \lambda^T Cx \ \ \underline{\text{subject to}}\ x \in X.$$

While the description of the efficient face in terms of the optimal set for the program Pλ enables us to recognize previously generated faces, it is important to note that this description is not gained at the cost of extra computational effort: the algorithm recognizes f(T,F) as efficient by constructing a tableau which immediately gives the values for a vector (v,w) ε G(F). As before the corresponding λ is simply v + e.

The precise statement of the main algorithm as given in [5] is complicated and requires some additional notation. We will try here to describe the main features of the algorithm by illustrating its use on an example having three objectives. We will not be able to illustrate all the various cases and features of the algorithm but this example will provide the reader with a basic understanding of how the algorithm works. More details can be found in [5] and [8].

4. AN ILLUSTRATIVE EXAMPLE. Consider the multiple objective problem

$$\underline{\max} \begin{pmatrix} 4 & 1 & 2 \\ 1 & 3 & -1 \\ -1 & 1 & 4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$$

$$\underline{\text{subject to}}$$

$$\begin{pmatrix} 1 & 1 & 1 \\ 2 & 2 & 1 \\ 1 & -1 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \leqq \begin{pmatrix} 4 \\ 6 \\ -1 \end{pmatrix} \quad \text{and } x \geqq 0 .$$

The feasible region X for this problem is given in Figure 1.

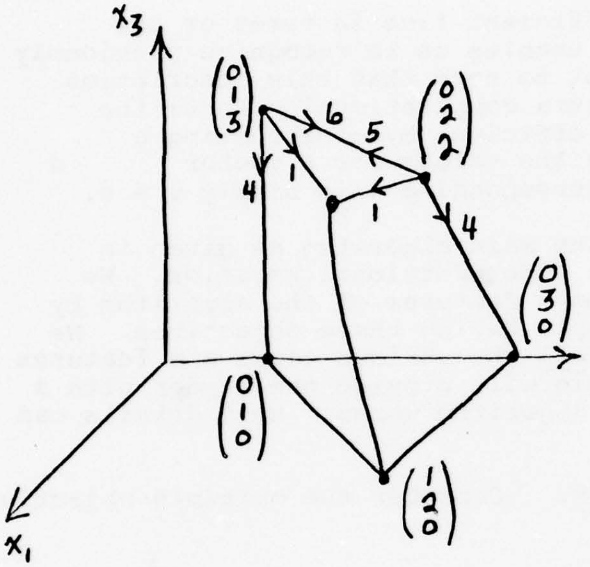

x_3
$\begin{pmatrix}0\\1\\3\end{pmatrix}$
6
5
$\begin{pmatrix}0\\2\\2\end{pmatrix}$
4
1
1
4
$\begin{pmatrix}0\\3\\0\end{pmatrix}$
x_2
$\begin{pmatrix}0\\1\\0\end{pmatrix}$
$\begin{pmatrix}1\\2\\0\end{pmatrix}$
x_1

Figure 1: The feasible region X

The reason for the labeled arrows at the vertices will become clear as we proceed. To initiate the method an efficient extreme point is required. The point $x^o = (0,1,3)^T$ is such that program Px^o of section 2 has an optimal value of zero. Thus $x^o \in E$ and will be used as our starting point. If we introduce slack variables x_4, x_5, and x_6 we obtain the tableau below which represents the problem

	x_1	x_2	x_3	x_4	x_5	x_6
0	-4	-1	-2	0	0	0
0	-1	-3	1	0	0	0
0	1	-1	-4	0	0	0
4	1	1	1	1	0	0
6	2	2	1	0	1	0
-1	1	-1	0	0	0	1

The nonbasic variables associated with the initial efficient vertex x^o are x_1, x_4 and x_6. Thus, two pivots on the above tableau yields the following tableau T_o.

T_o

	x_1	x_2	x_3	x_4	x_5	x_6
7	-1	0	0	2	0	1
0	-6	0	0	-1	0	-4
13	8	0	0	4	0	3
3	2	0	1	1	0	1
1	2	0	0	-1	1	1
1	-1	1	0	0	0	-1

$N_{T_o} = \{1,4,6\}$.

To determine which of the nonbasic indices N_{T_o} correspond to efficient edges incident to x_o we form the tableau for the linear system S_{T_o} as indicated by the corollary to Theorem 3,

S_{T_o}

	v_1	v_2	v_3	w_1	w_4	w_6
1	1	6	-8	1	0	0
5	-2	1	-4	0	1	0
0	-1	4	-3	0	0	1

Notice that this tableau also represents the set $G(\emptyset)$. By inspection, we see that the basic feasible solution associated with S_{T_o} has $w_6 = 0$ and also that w_1 can be made nonbasic in a single pivot. Thus, w_1 and w_6 are clearly nonredundant in $G(\emptyset)$. Attempting to minimize w_4 over S_{T_o} gives $w_4 \geqq 5$. Therefore, increasing the nonbasic variable x_4 in T_o does not lead along an efficient edge while increasing x_1 or x_6 does yield an efficient edge. Notice the arrows in Figure 1 indicate the respective edges for increasing the nonbasic variables.

Thus, here we have $J_{T_o} = \{1,6\}$. We will now show how subroutine FACE uses this set to construct maximal efficient faces incident x^o. First, we pick an index in J_{T_o}, say $j = 6$. Then we find a tableau for $G(\emptyset)$ with w_6 nonbasic at level zero.

$$\begin{array}{|c|cccccc|} & & v & & w_1 & w_4 & w_6 \\ \hline 1 & 5/2 & 0 & -7/2 & 1 & 0 & -3/2 \\ 5 & -7/4 & 0 & -13/4 & 0 & 1 & -1/4 \\ 0 & -1/4 & 1 & -3/4 & 0 & 0 & 1/4 \\ \hline \end{array} \qquad F = \{6\},\ G(F) \neq \emptyset\ .$$

This tableau shows that $G(\{6\}) \neq \emptyset$. With w_6 <u>blocked at zero</u> (kept nonbasic) we now proceed to check if we can find a solution with $w_1 = 0$ as well. In one pivot we obtain

$$\begin{array}{|c|cccccc|} & & v & & w_1 & w_4 & w_6 \\ \hline .4 & 1 & 0 & -1.4 & .4 & 0 & -.6 \\ 5.7 & 0 & 0 & -5.7 & .7 & 1 & -1.3 \\ .1 & 0 & 1 & -1.1 & .1 & 0 & .1 \\ \hline \end{array} \qquad F = \{1,6\},\ G(F) \neq \emptyset.$$

Thus in Figure 1, the face defined by increasing both x_1 and x_6 must be efficient according to Theorem 3 and in fact must be a maximal efficient face. Notice that with w_1 and w_6 both blocked at zero, we cannot decrease w_4 below 5.7. Thus, the first maximal efficient face is defined by

$f(T_o,F)$ where $F = \{1,6\}$.

So far the variables v have not played much of a role. In the final tableau above the values of the components of v are crucial as shown by the following theorem.

<u>Theorem</u> 4. If $(\bar{v},\bar{w})$ is the basic feasible solution for the tableau for $G(\emptyset)$ that identifies a maximal efficient face $f(T,F)$, then $f(T,F)$ is identical to the set of optimal solutions to the linear program

$$P\bar{\lambda} : \underline{\max}\ \bar{\lambda}^T Cx \quad \text{subject to } x \in X$$

$$\text{where } \bar{\lambda} = \bar{v} + e.$$

<u>Proof</u>. see [5].

In our example, the face $f(T_o, \{1,6\})$ is associated with the weighting vector $\lambda^0 = (1.4, 1.1, 1)^T$. Notice that $(\lambda^o)^T C$ is normal to the face $f(T_o, \{1,6\})$ of Figure 1. Having found all maximal efficient faces incident to x^o we then pivot on tableau T_o to an adjacent efficient vertex, say $x^1 = (0,2,2)^T$, see Figure 1. This yields the tableau

$$T_1 \quad \begin{array}{c|cccccc} & x_1 & x_2 & x_3 & x_4 & x_5 & x_6 \\ \hline 5 & -3 & 0 & 0 & 3 & -1 & 0 \\ 1 & 2 & 0 & 0 & -5 & 4 & 0 \\ 9 & 2 & 0 & 0 & 7 & -3 & 0 \\ \hline 2 & 0 & 0 & 1 & 2 & -1 & 0 \\ 1 & 2 & 0 & 0 & -1 & 1 & 1 \\ 2 & 1 & 1 & 0 & -1 & 1 & 0 \end{array} \quad ; \ N_{T_1} = \{1,4,5\} \ .$$

We now proceed to find those faces incident to x^1 that are maximal efficient. Notice that because of the pivot, we know that $f(T_o, \{1,6\}) = f(T_1, \{1,5\})$ since indices 6 and 5 were exchanged as nonbasic indices in the pivot, see the arrows in Figure 1. In particular, this allows us to conclude that J_{T_1} contains $\{1,5\}$. That is, pivoting in the x_1 or x_5 column of T_1 corresponds to going along an efficient edge. We need only check to see if $4 \in J_{T_1}$ and we find by examining S_{T_1} that $J_{T_1} = \{1,4,5\}$.

By careful bookkeeping procedures we can make good use of previous work. For example, notice that it is not possible for $F = \{5,4\}$ to identify an efficient face since at the previous tableau T_o the set $\{6,4\}$ did not define an efficient face (and 6 replaced 5 on the pivot). The only subset of J_{T_1} we need to check is $\{1,4\}$, and after obtaining S_{T_1} two pivots yield

$$S_{T_1} \quad \begin{array}{c|cccccc} & & & & w_1 & w_4 & w_5 \\ \hline 5/3 & 1 & 0 & -8/3 & 5/9 & 2/9 & 0 \\ 2 & 0 & 1 & -3 & 1/3 & 1/3 & 0 \\ 19/3 & 0 & 0 & -19/3 & 7/9 & 10/9 & 1 \end{array} \quad .$$

Thus, $f(T_1, \{1,4\})$ is a maximal efficient face as well as $f(T_1, \{1,5\})$. Using Theorem 4, we see that $f(T_1, \{1,4\})$ is identical to the set of optimal solutions to $P\lambda'$ where $\lambda' = (8/3, 3, 1)^T$.

In this example it turns out that there are only two maximal efficient faces; namely the ones associated with λ^o and λ^1. Of course, the algorithm continues to pivot to unvisited efficient vertices and continues to examine incident faces. If a new vertex visited is not adjacent to the previous vertex, it is crucial to know whether or not that vertex is on a previously generated maximal efficient face. Given a tableau T, let

$$A_T = \{(F,\ell) \mid (-\lambda^{\ell})^T C \geqq 0 \text{ and } F=\{j \mid (\lambda^{\ell})^T C^j = 0\}\} .$$

Having identified the weighting vectors λ^{ℓ} for previous faces, A_T simply shows how to describe (relative to T) those previously generated maximal efficient faces containing x_T. Similarly, the new vertex may be adjacent to the previous vertex but may be in a previously generated maximal face not containing the previous vertex. Using a set similar to A_T, these faces can also be recognized by using the weighting vectors λ^{ℓ}.

Upon completion, the algorithm provides a list of all efficient extreme points, a list of all maximal efficient faces of X given as the convex hulls of sets of efficient vertices, and finally an implicit description of E as the union of optimal sets for the linear programs $P\lambda^{\ell}$, $\ell = 1,2,\ldots,L$ where L is the number of maximal efficient faces.

References

1. COCHRANE, JAMES L. and ZELENY, MILAN, eds., Multiple Criteria Decision Making, U. of S. C. Press, Columbia, S.C., 1973.

2. ECKER, J.G. and KOUADA, I.A., "Finding Efficient Points for Multiple Objective Linear Programs", Mathematical Programming 8 (1975) 375-377.

3. ECKER, J.G. and KOUADA, I.A., "Finding All Efficient Extreme Points for Multiple Objective Linear Programs", Mathematical Programming 14 (1978) 249-261.

4. ECKER, J.G. and HEGNER, NANCY S., "On Computing an Initial Efficient Extreme Point", Operational Research Quarterly, to appear.

5. ECKER, J.G., HEGNER, NANCY S., and KOUADA, I.A., "Generating All Maximal Efficient Faces for Multiple Objective Linear Programs", Journal of Optimization Theory and Application, to appear.

6. EVANS, J.P. and STEUER, R.E., "A Revised Simplex Method for Linear Multiple Objective Programs", Math. Prog. 5 (1973) 54-72.

7. GAL, TOMAS, "A General Method for Determining the Set of All Efficient Solutions to a Linear Vector-maximum Problem", European Journal of Operations Research 1 (1977) 307-322.

8. HEGNER, NANCY S., "Multiple Objective Linear Programming", Dissertation, Rensselaer Polytechnic Institute, Troy, New York (1977).

9. ISERMANN, H., "Proper Efficiency and the Linear Vector Maximum Problem", OR 22 (1974) 189-191.

10. ISERMANN, H., "The Enumeration of the Set of All Efficient Solutions for a Linear Multiple Objective Program", Operations Research Quarterly, 28 (1977).

11. PHILIP, JOHAN, "Algorithms for the Vector Maximization Problem", Math. Prog. 2 (1972) 207-229.

12. STARR, MARTIN K. and ZELENY, MILAN, eds., Multiple Criteria Decision Making, North Holland Publishing Co., New York, N.Y., (1977).

13. YU, P.L. and ZELENY, M., "The Set of All Nondominated Solutions in Linear Cases and a Multicriteria Simplex Method", J. Math. Anal. and App. 49 (1975) 430-468.

14. ZIONTS, S., ed., Multiple Criteria Problem Solving, Springer-Verlag, New York, (1978).

APPLICATION OF JENSEN'S INEQUALITY FOR ADAPTIVE SUBOPTIMAL DESIGN

Chelsea C. White, III*

and

David P. Harrington

School of Engineering and Applied Sciences
University of Virginia
Charlottesville, Va. 22901

SUMMARY

In this paper it is shown that if the expected cost-to-go functions generated by a suboptimal design for a partially observed (discrete time) Markov decision problem with a specific state measurement quality are concave, then the suboptimal design has a desirable adaptivity characteristic relative to that state measurement quality. Optimal strategies are shown to possess this adaptivity characteristic, as does a suboptimal design presented in an example.

* The research of the first author has been supported by NSF Grant 76-15774.

1. INTRODUCTION AND NOTATION

A desirable feature for a suboptimal design to possess is that if state observation quality improves, then so does the performance of the suboptimal design. (The reason for seeking ways to improve the quality of state measurements is precisely to improve system performance.) This paper presents sufficient conditions for a control law, or strategy, to have such an adaptivity characteristic relative to a given observation quality. These conditions are the concavity of the expected cost-to-go functions generated by the strategy and the given observation quality. It is shown that under quite general conditions, optimal strategies possess this adaptivity feature, providing an alternate proof to the fact that the open-loop feedback controller is superior to the open-loop controller. An example illustrates a strictly suboptimal design having such a characteristic.

In this section we define the problem to be examined. Preliminary results are then presented in Section 2. An important lemma is stated whose proof is based on Jensen's inequality. A precise definition of adaptivity relative to a specific observation quality is given prior to the presentation of the main result in Section 3. This definition captures much of the intent of the definition found in Bertsekas (1976). (In-depth discussions of the myriad definitions of adaptivity can be found in Asher, Andrisani, and Dorato (1976) and Saridis (1977).) A suboptimal design having the adaptivity characteristic is presented in an example.

A partially observed Markov decision problem (POMDP) is defined in terms of the set $(S,A,Z,\beta,r,r_0,p,q_0,q,N)$. S, A, and Z are the respective state spaces for the state process $\{s_t:t=0,1,\ldots,N\}$, the action process $\{a_t:t=0,1,\ldots,N-1\}$, and the observation process $\{z_t, t=0,\ldots,N-1\}$. S, A and Z will always be taken to be complete separable metric spaces, and $\mathcal{B}_S$, $\mathcal{B}_A$, and $\mathcal{B}_Z$ will be used to denote the collection of Borel sets of these spaces. β is a discount factor, and we will always take $\beta\varepsilon(0,1]$. The one-stage cost function r and the terminal cost function r_o will be taken to be lower semianalytic functions on $S \times A$ and S, respectively. The definition of a semianalytic set may be found, for example, in Shreve (1977). The state process s_t makes transitions according to the state transition kernel $p=p(\cdot|s,a)$, which is taken to be a stochastic kernel on $\mathcal{B}_S\times S\times A$, i.e., for fixed $(s,a)\varepsilon S\times A$, $p(\cdot|s,a)$ is a probability measure on $\mathcal{B}_S$, and for fixed $B\varepsilon\mathcal{B}_S$, $p(B|\cdot)$ is a Borel measurable real valued function on $S\times A$. The initial observation kernel and observation kernel are denoted by q_o and q , and will be taken to be stochastic kernels on $\mathcal{B}_S\times S$ and $\mathcal{B}_S\times S\times A$, respectively. They satisfy the two relationships

$$P(z_o\varepsilon B_o|s_o)=\int_{B_o} q_o(dz|s_o)$$

and

$$P(z_t\varepsilon B_1|s_t,a_t)=\int_{B_1} q(dz|s_t,a_t)$$

for all $B_o,B_1\varepsilon\mathcal{B}_S$. The integer N, the length of the observation and control period, is called the horizon, and will always be such that $1\leq N\leq\infty$.

Let Y represent the space of probability measures on S. There exists a natural topology on Y making Y a separable metric space (cf. Theorem 6.1, Parthasarathy (1967)). An admissible strategy for the POMDP will be a sequence

$\pi=(u_0,\ldots,u_{N-1})$ such that for $0\leq t\leq N-1$, the policy $u_t(a_t|\bar{y},i_t)$ is a universally measurable stochastic kernel on $\mathcal{B}_A\times Y\times Z\times(A\times Z)^t$. The term $i_t\varepsilon Z\times(A\times Z)^t$ represents the history of the observations and actions taken up to time t-1 and the observation at time t, and $\bar{y}$ is the initial distribution on S.

The object of the POMDP is to choose an admissible strategy so as to minimize the quantity

$$E_{\bar{y}}\{\sum_{t=0}^{N-1}\beta^t r(s_t,a_t)+\beta^N r_o(s_N)\},$$

where $E_{\bar{y}}$ denotes expectation with respect to the initial distribution $\bar{y}\varepsilon Y$ on S.

Conditions have been determined (and are presented, for example, in Shreve (1977) and Striebel (1975)) which insure the existence of ε-optimal and optimal nonrandomized strategies that are dependent on $(\bar{y},i_t)$ only through the measure $y_t\varepsilon Y$, where $y_t(C)=P(s_t\varepsilon C|\bar{y},i_t)$, $C\varepsilon\mathcal{B}_S$. These conditions, and the fact that it is extremely difficult to justify the use of randomized strategies in applications, serve as justification that throughout the remainder of this paper, consideration will only be given to strategies composed of nonrandom policies dependent on $(\bar{y},i_t)$ only through y_t. The set of all such strategies will be denoted by Π.

The intent of this paper is to present sufficient conditions for a (not necessarily optimal) strategy $\pi\varepsilon\Pi$ to improve expected performance given improved observation quality. These conditions are presented following several results and definitions.

2. PRELIMINARY RESULTS

Let $\sigma(\cdot|y,a)$ and $\lambda(\cdot|z,y,a)$ be stochastic kernels defined on $\mathcal{B}_Z\times Y\times A$ and $\mathcal{B}_S\times Z\times Y\times A$ which satisfy

$$\sigma(B|y,a)=P(z_{t+1}\varepsilon B|y_t=y,a_t=a)$$
$$=\int_B \sigma(dz|y,a)$$

and

$$\lambda(C|z,y,a)=P(s_{t+1}\varepsilon C|z_{t+1}=z,y_t=y,a_t=a)$$
$$=\int_C \lambda(ds|z,y,a).$$

(The proof that such kernels exist may be found in Shreve (1977).)

If $J_N^{\pi q}(\bar{y})$ is the expected cost to be accrued over horizon N by strategy $\pi\varepsilon\Pi$ under initial distribution $\bar{y}$ and observation kernel q, then $J_N^{\pi q}$ can be shown to satisfy the following recursive equation (RE) and boundary condition:

$$J_t^{\pi q}(y)=R(y,a')+\beta\int_Z J_{t-1}^{\pi q}[\lambda(\cdot|z,y,a')]\sigma(dz|y,a'),$$

$$J_0{}^{\pi q}(y)=R_0(y),$$

where $a'=u_{N-t}(y),R_0(y)=\int_S r_0(s)y(ds)$ and $R(y,a)=\int_S r(s,a)y(ds)$.

When N=∞, the RE is identical to that above, except that $J_t^{\pi q}$ and $J_{t-1}^{\pi q}$ are replaced by $\lim_{t\to\infty} J_t{}^{\pi q}$ in situations sufficiently regular to insure that the limit exists and may be interchanged with the integral. Such regularity conditions can be found in Striebel (1975).

We now present a definition of the relative worth of state measurements taken by two observation kernels and two important lemmas.

$$G(F) = \{(v,w) \geqq 0 \mid C^T v + w = -C^T e,\ w_j = 0 \text{ for } j \in F\} \quad ,$$

DEFINITION 2.1. The observation kernel q is said to represent improved measurement quality compared to observation kernel q', i.e., q'<q, if for each action $a \epsilon A$, there exists a stochastic kernel γ on $\mathcal{B}_Z \times Z \times A$ such that

$$q'(B|s,a) = \int_Z \gamma(B|z,a)q(dz|s,a)$$

for all $s \epsilon S$, $z \epsilon Z$, and $B \epsilon \mathcal{B}_Z$.

This definition is intuitively reasonable; the quality of the information received by the channel associated with q' is equivalent to the quality of information received by the channels associated with q and γ placed in series.

Note that for the finite state case, q' and q are equivalent to stochastic matrices, and q'<q if and only if there exists a stochastic matrix γ such that $q\gamma = q'$. The relative information quality of linear systems having additive Gaussian noise has the following simple characterization. Let $z = As + \xi$, $z' = As + \xi'$, where $\xi \sim N(0,\Sigma)$, $\xi' \sim N(0,\Sigma')$, and Σ and Σ' are both covariance matrices. Let $\Sigma' \leq \Sigma$ if $\Sigma' - \Sigma$ is positive semi-definite. It is easily shown that q'<q, where q and q' may now be thought of as Gaussian densities, if $\Sigma' \leq \Sigma$ by letting the density $\gamma(z'|z)$ be associated with a normally distributed random variable with mean z and covariance $\Sigma' - \Sigma$ which is independent of ξ.

Let σ' and λ' be defined as were σ and λ except that q is replaced by q'. A few technical preliminaries are necessary before we proceed. Although we have indicated that many of the stochastic kernels used below depend on a_t, this dependence will be suppressed throughout the describtion; it is described as a face f(T,F) incident to some efficient vertex x_T and it is also described by identifying a

remainder of this section to simplify some notation.

If $\sigma'(B|y)=0$, for $B\varepsilon\mathcal{B}_Z$ then we must have $\gamma(B|z)=0$ for a set of values z which have probability one with respect to the measure $\sigma(\cdot|y)$. Since the Borel sets in Z are countably generated, there exists a set $D_y\varepsilon\mathcal{B}_Z$ such that $\sigma(D_y|y)=0$ and $\sigma(B|y)=0$ implies $\{z:\gamma(B|z)>0\}\subset D_y$. Thus, for almost all z, $[\sigma(\cdot|y)]$, the measure $\gamma(\cdot|z)$ is absolutely continuous with respect to $\sigma'(\cdot|y)$, and the Radon-Nikodym derivative $\frac{d\gamma(\cdot|z)}{d\sigma'(\cdot|y)}$ exists and is finite a.s. $[\sigma(\cdot|y)]$.

Recall that $\lambda(C|z,y)=P(s_{t+1}\varepsilon C|z_{t+1}=z,y_t=y)$. Define the stochastic kernel $\mu(\cdot,\cdot|y)$ on $\mathcal{B}_Z\times\mathcal{B}_S\times Y\times A$ by $\int_{B\times C}\mu(dz,ds|y)=P(z_{t+1}\varepsilon B,s_{t+1}\varepsilon C|y_t=y)$. It is easy to show that for fixed $C\varepsilon\mathcal{B}_S$, μ is a substochastic kernel on $B_Z\times Y\times A$ that is absolutely continuous with respect to $\sigma(\cdot|y)$. Since $\int_B\lambda(C|z,y)\,\sigma(dz|y)=\mu(B\times C|y)$ for $B\varepsilon\mathcal{B}_Z$, we must have $\lambda(C|z,y)=\frac{d\mu(\cdot,C|y)}{d\sigma(\cdot|y)}$ a.s. $[\sigma(\cdot|y)]$.

LEMMA 2.1. Define the measure $\Lambda(\cdot|z,y)$ on $\mathcal{B}_Z\times Z\times Y\times A$ by

$$\Lambda(B|z',y)=\frac{d\gamma(\cdot|z')}{d\sigma'(\cdot|y)}\,\sigma(B|y).$$

Then

$$\lambda'(C|z',y)=\int_Z\Lambda(dz|z',y)\lambda(C|z,y).$$

Proof: Let $B'\varepsilon\mathcal{B}_Z$, and let dz' in the kernel σ' denote integration over B'.

Since

$$\int_{B'}\int_Z\Lambda(dz|z',y)\lambda(C|z,y)\sigma'(dz'|y)$$

$$=\int_{B'}\int_Z\frac{d\gamma(\cdot|z')}{d\sigma'(\cdot|y)}\cdot\frac{d\mu(\cdot,C|y)}{d\sigma(\cdot|y)}\,\sigma(dz|y)\sigma'(dz'|y)$$

$$\int_{B'}\int_Z\gamma(dz|z')\mu(dz',C|y)$$

$$=P(z_{t+1}\varepsilon B',s_{t+1}\varepsilon C|y_t=y),$$

under information quality q', it follows that $\int_Z \Lambda(dz|z',y)\lambda(C|z,y)$ satisfies the defining relation for the Radon-Nikodym derivative λ'. □

LEMMA 2.2. Let $q'<q$. Then for any concave function $g:Y\to R$,

$$\int_Z g[\lambda(\cdot|y,z)]\sigma(dz|y)\le\int_Z g[\lambda'(\cdot|y,z')]\sigma'(dz'|y)$$

for all $y\in Y$.

Proof: From Jensen's inequality (cf. Chung (1968), p. 45) and Lemma 2.1

$$g[\lambda'(B|y,z')]\ge\int_Z\Lambda(dz|z',y)\; g[\lambda(B|y,z)]$$

for all $y\in Y$, $z\in Z$ and $B\in\mathcal{B}_Z$. Fubini's Theorem and the definition of Λ then imply that

$$\int_Z g[\lambda'(B|y,z')]\sigma'(dz'|y)$$

$$\ge \int\int_{Z\times Z}\Lambda(dz|z',y)g[\lambda(B|y,z)]\sigma'(dz'|y)$$

$$=\int_Z\gamma(dz'|z)\int_Z g[\lambda(B|y,z)]\sigma(dz|y)$$

$$=\int_Z g[\lambda(y,z)]\sigma(dz|y).$$

□

3. MAIN RESULTS

Conditions are now presented on the pair (π,q) which imply that if $q<q'$ then $J_N^{\pi q'}\le J_N^{\pi q}$ for $N<\infty$ and for $N=+\infty$ when $\lim_{N\to\infty} J_N^{\pi q}$ exists and satisfies the RE.

DEFINITION 3.1. A strategy $\pi\in\Pi$ is q-adaptive if for all q' such that $q<q'$, $J_N^{\pi q'}\le J_N^{\pi q}$.

THEOREM 3.1. The strategy $\pi\in\Pi$ is q-adaptive if $J_t^{\pi q}$ is concave on Y for all $t=1,\dots,N$.

The nonbasic variables associated with the initial efficient

Proof: Assume $q<q'$. Trivially, $J_k^{\pi q'} \le J_k^{\pi q}$ on Y for k=0; assume this inequality holds for k=t-1. Then using Lemma 2.2 and letting $a=u_{N-t}(y)$,

$$J_t^{\pi q'}(y)=R(y,a)+\beta\int_Z J_{t-1}^{\pi q'}[\lambda'(\cdot|y,z,a)]\sigma'(dz|y,a)$$

$$\le R(y,a)+\beta\int_Z J_{t-1}^{\pi q}[\lambda'(\cdot|y,z,a)]\sigma'(dz|y,a)$$

$$\le R(y,a)+\beta\int_Z J_{t-1}^{\pi q}[\lambda(\cdot|y,z,a)]\sigma(dz|y,a)$$

$$=J_t^{\pi q}(y).$$

The result follows by induction. □

Observe that the proof of the *Theorem* implies also that if measurement quality degrades for a q-adaptive strategy then so does the performance of the strategy.

We assume that conditions are satisfied (such as those presented in Striebel (1975) and Shreve (1977)) which imply that if $u_{N-t}(.)$ minimizes the right hand side of the RE over A for all t=1,...,N, then $\pi=(u_0,\ldots,u_{N-1})$ is an optimal strategy. The next result is required prior to proving the following corollary. The proof given is a more general version of a proof given by Åström (1969) for the case where both the state and observation spaces are finite. Dependence on a_t will be suppressed in the statement of the corollary and its proof.

LEMMA 3.1. Let g by a real valued concave function on Y. Then $h(y)=\int_Z g[\lambda(\cdot|y,z)]\sigma(dz|y)$ is a concave function of y.

Proof: Suppose that $y=\alpha y_1+(1-\alpha)y_2$. Since

$$\sigma(B|y)=\sigma(B|\alpha y_1+(1-\alpha)y_2)$$
$$=\alpha\sigma(B|y_1)+(1-\alpha)\sigma(B|y_2),$$

we must have $\sigma(B|y)=0 \Rightarrow \sigma(B|y_1)=\sigma(B|y_2)=0$. Thus $\sigma(\cdot|y_i)$ is absolutely continuous with respect to $\sigma(\cdot|y)$ for i=1,2. Let $d_i = \frac{d\sigma(\cdot|y_i)}{d\sigma(\cdot|y)}$ be the Radon-Nikodym derivative of $\sigma(\cdot|y_i)$ with respect to $\sigma(\cdot|y)$ for i=1,2. Using the defining relationship for Radon-Nikodym derivatives, it is easily shown that

$$\lambda(\cdot|y,z)=\delta\lambda(\cdot|y_1,z)+(1-\delta)\lambda(\cdot|y_2,z)$$

where $\delta=\alpha d_1$ and $1-\delta=(1-\alpha)d_2$.

Thus

$$\int_Z g[\lambda(\cdot|y,z)]\sigma(dz|y)$$

$$\geq\int_Z\{\delta g[\lambda(\cdot|y_1,z)]+(1-\delta)g[\lambda(\cdot|y_2,z)]\}\sigma(dz|y)$$

$$=\alpha\int_Z g[\lambda(\cdot|y_1,z)]\sigma(dz|y_1)$$

$$+(1-\alpha)\int_Z g[\lambda(\cdot|y_2,z)\sigma(dz|y_2)] \qquad \square$$

The following result guarantees the existence of q-adaptive controllers when optimal controllers in Π exist.

DEFINITION 3.2. A strategy $\pi\varepsilon\Pi$ is said to be q-optimal if it is such that $J_N^{\pi q}\leq J_N^{\pi' q}$ for all $\pi'\varepsilon\Pi$.

COROLLARY 3.1. If π is q-optimal, then π is q-adaptive.

Proof: We wish to show that $J_t^{\pi q}$ is concave for all t when π is q-optimal. $J_o^{\pi q}$ is concave; assume $J_{t-1}^{\pi q}$ is concave. By Lemma 3,

$$\int_Z J_{t-1}^{\pi q}[\lambda\ (\cdot|y,z,a)]\sigma\ (dz|y,a)$$

is concave on Y for each $a\varepsilon A$. The concavity of $J_t^{\pi q}$ then follows from the

T_o to an adjacent efficient vertex, say
Figure 1. This yields the tableau

concavity of R(y,a) on Y for each a, the fact that sums of concave functions are concave, and the fact that the infimum of concave functions is concave. □

In the context of our notation, a prominent definition of adaptivity is as follows: π is adaptive if $J_N^{\pi q} \leq J_N^{\pi' q'}$, where $q' < q$, q' is independent of the state (i.e. complete unobservability), and π' is q'-optimal, cf. Bertsekas (1976). The strategy π' is often referred to as the open-loop controller (OLC). A common suboptimal design is to apply the OLC with an observation kernel q, where (of course) $q' < q$, i.e., $\pi = \pi'$. Such a suboptimal design is called the open-loop feedback controller (OLFC). A proof that the OLFC is superior to the OLC can be found in Bertsekas (1976); we note that this result follows directly from Corollary 3.1.

The following corollary generalizes results in Åström (1965) and White (1976) under the assumption that an optimal strategy exists.

COROLLARY 3.2. Let π and π' be q-optimal and q'-optimal, respectively, where $q' < q$. Then, $J_t^{\pi q} \leq J_t^{\pi' q'}$ for all t.

Proof: Corollary 1 and the Theorem imply that $J_t^{\pi' q} \leq J_t^{\pi' q'}$ for all t. The q-optimality of π implies $J_t^{\pi q} \leq J_t^{\pi' q}$ for all t. □

EXAMPLE [Sternby (1976)] We now present a strictly suboptimal q-adaptive design for a finite state POMDP that is based on the solution of its associated completely observed problem. Consider the following problem (described more fully by Sternby):

cases, the algorithm continues to pivot to unvisited efficient vertices and continues to examine incident faces. If a new vertex visited is not adjacent to the previous vertex, it is

$$r(a) = \begin{bmatrix} p_1 + p_5 \\ p_1 + p_5 \\ p_2 + p_6 \\ p_2 + p_6 \end{bmatrix} \qquad P(a) = \begin{bmatrix} p_1 & p_3 & 0 & p_5 \\ p_1 & p_3 & 0 & p_5 \\ p_2 & 0 & p_4 & p_6 \\ p_2 & 0 & p_4 & p_6 \end{bmatrix}$$

$$r_o = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \qquad q = \begin{bmatrix} 0 & 1 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$$

where $A=\{a:0\leq a\leq 1\}$, $P(a)$ is the state transition matrix, and each of the $p_i=p_i(a)$, $i=1,\ldots,6$, is defined as in Sternby (1976). It is straightforward to show that the optimal strategy for the completely observed undiscounted N-horizon ($N\leq\infty$) problem is stationary (hence myopic) and is composed of the policy $\delta^*(1)=\delta^*(2)=0.2$ and $\delta^*(3)=\delta^*(4)=0.8$. Let $b_t=P(s_t\varepsilon\{1,2\}|(\bar{y},i_t))$, and consider the stationary suboptimal strategy π associated with the policy $\delta(b)=0.8$ if $b\leq 1/2$ and $\delta(b)=0.2$ if $b>1/2$, which is equivalent to taking the most probable state as the state estimate and is an example of a certainty-equivalent controller. Assume the observation quality associated with π is q as defined above. Then, as shown in Sternby (1976), $J_t^{\pi q}(b)$ is concave for all t and hence π is q-adaptive. From Figure 4 in Sternby's paper, note that π is superior to the OLFC for $N=\infty$; hence, a suboptimal design based on the easily computable solution of a completely observed problem is both in some sense adaptive (specifically, q-adaptive), superior

Efficient Solutions to a Linear Vector-maximum Problem",
European Journal of Operations Research 1 (1977) 307-322.

to the OLFC, and easily implementable. The suboptimal strategy is passively adaptive in that it makes no effort to identify the system state, as is also true of the OLFC (Bar-Shalom and Tse (1974)).

REFERENCES

Asher, R. B., Andrisani, D. and Dorato, P. (1976). Bibliography on adaptive control systems. Proc. IEEE 64 1226-1240.

Åström, K. J. (1965). Optimal control of Markov processes with incomplete state information. J. Math. Analysis. Appl. 10 174-205.

Åström, K. J. (1969). Optimal control of Markov processes with incomplete state information II. J. Math. Analysis Appl. 25 403-406.

Bar-Shalom, Y. and Tse, E. (1974). Dual effect, certainty equivalence, and separation in stochastic control. IEEE Trans. Aut. Cont. AC-19 494-500.

Bertsekas, D. P. (1976). Dynamic Programming and Stochastic Control. Academic Press, New York.

Chung, K. L. (1968). A Course in Probability Theory. Harcourt, Brace and World, Inc., New York.

Parthasarathy, K. R. (1967). Probability Measures on Metric Spaces. Academic Press, New York.

Saridis, G. N. (1977). Self-Organizing Control of Stochastic Systems. Marcel Dekker, New York.

Shreve, S. E. (1977). Dynamic programming in complete separable spaces. Ph.D. dissertation, Department of Mathematics, University of Illinois.

Striebel, C. (1975). Optimal Control of Discrete Time Stochastic Systems. Springer-Verlag, Berlin.

Sternby, J. (1976). A simple dual control problem with an analytical solution. IEEE. Trans. on Aut. Cont. AC-21 840-844.

White, C. C. (1976). Application of two inequality results for concave functions to a stochastic optimization problem. J. Math. Analysis Appl. 55 347-350.

PROBLEMS WITH SOFTWARE DEVELOPMENT IN THE SOVIET UNION

S. E. Goodman

Science Division
U. S. Army Foreign Science and Technology Center

Department of Mathematics and the Woodrow Wilson School
of Public and International Affairs
Princeton University

Department of Applied Mathematics and Computer Science
and the Center for Russian and East European Studies
University of Virginia

ABSTRACT. The development of computing in the USSR has tended to follow the Western technical pattern, but only within the last decade have the Soviets committed themselves to the production and installation of complex general purpose computer systems in large enough numbers to make advances in software essential. However, there are few shortcuts to the attainment of a national software capacity on a level anything like that which exists in the United States. The USSR does have considerable resources for the pursuit of this goal, yet the task is so large and complex, and their economic system is so ill structured to support many of the practices that have worked well for us, that Soviet overall progress will be slow, painful and evolutionary.

I am a consultant for the U. S. Army Foreign Science and Technology Center on the subject of Soviet computing. This brief introduction to some of the problems that the USSR is having with software development is part of a more comprehensive study that I am doing for FSTC [4,5]. What I will say here during this short talk must be recognized as an oversimplified description of a complex subject, and that the views expressed here do not necessarily reflect official opinion or policy of the US Army or any other branch of the US Government.

I should also make it clear that I will be talking about the civilian side of Soviet software. The military side is probably doing better, although there are reasons to believe that software development for the military suffers to some degree from most or all of the difficulties that I will mention.

A few statistics on the US software industry will give you some idea of the scope of the technology. There are over 500,000 professional programmers and systems analysts in the US and perhaps twice as many software-related people with lesser skills (e.g. keypunch operators, tape librarians). In terms of their basic salaries and material support alone, software is a $20-$30 billion per year industry. It is much more difficult to ascertain the real value of the industry. For example, these figures do not count the millions of business men, government bureaucrats, scientists and engineers who spend some of their professional time programming.

Hardware-software cost trends for the US are shown in Figure 1 [2].* As a result of 30 years of widespread, painful experience, we have gotten to the point where we have learned to produce and to effectively use a tremendous variety of software products, many of which contain hundreds of thousands or millions of instructions. Clearly, any attempt by the USSR to develop a national software capability comparable to that which exists in the US would be a serious test of the strength and sophistication of the entire economic system.

It will be convenient to divide the factors that affect Soviet software development into four, somewhat overlapping, categories:

(1) Those that are related to priorities in the allocation of effort and other resources.

(2) Those that are dependent on hardware availability.

(3) Those that are dependent on the nature of Soviet institutions and economic practices.

and (4) Those that involve transfers from foreign sources.

In the time available, I can only say a little bit about each of the four categories.

Before the mid-1960s the Soviets made little effort to produce large quantities of suitable hardware and software intended for widespread general purpose use. No great need for this was perceived anywhere in the industrial hierarchy, and the cost would have been a great strain on their limited capabilities that would have been out of proportion to the benefits. It is moot to speculate as to what they could have done had they tried harder. The Soviets have had a number of successful high technology priority projects, and research and development for the early CPU (central processing unit) hardware they did produce was often of high quality considering the available circuitry. Nevertheless, it is doubtful if they could have matched the IBM S/360 system before the end of the 60s without an effort demanding an unreasonable commitment of resources. Until the early 60s the military and scientific/engineering communities were the only influential customers with an interest in computing. However both were less enamoured with computers than their American counterparts, and the Soviet industry developed only to the extent where it could respond to this relatively limited demand.

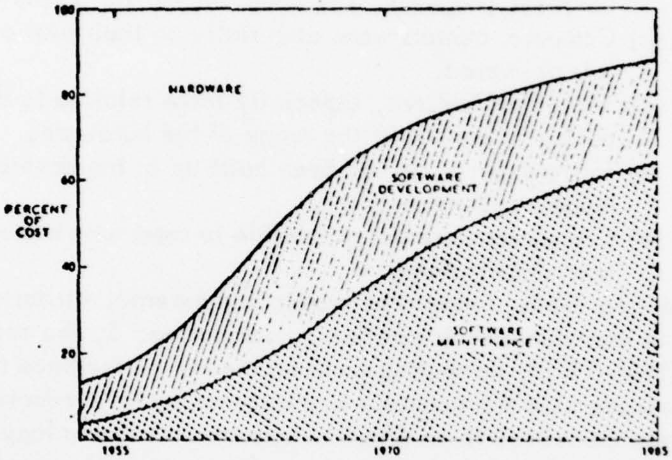
This rational, but shortsighted, policy crippled the development of hardware in the USSR. Of several major handicaps, three were particularly important.

(1) Soviet computers had small primary memories. Most machines were provided with no more, and frequently much less, than 32K words of poor quality core memory.

(2) For all practical purposes, disk storage was not widely available until 1973. Secondary storage was generally on poor quality magnetic tape units.

(3) There was a lack of suitable and reliable peripherals. Card readers and alphanumeric printers were not generally available until the

*Software maintenance refers to the location and correction of errors, adaptation of old systems to new hardware or software environments, and to the enhancement of existing systems.



Hardware-software cost trends.

Figure 1

mid-1960s. The units that were later produced, and their associated paper products, were of notoriously poor quality and reliability.

This situation was made worse by hardware vendor practices. They delivered nearly empty machines. Users had to write all but the most basic utility programs. Furthermore, the users had to maintain the hardware themselves. This eventually led to local engineering modifications that made it difficult or impossible for users with the same basic computer model to share software.

Because of these and other factors, the pre-Ryad software situation could be summarized as follows:

(a) Software existed in the form of many isolated pockets of machine language programs. There was very little portability.

(b) Computer centers were essentially on their own once the hardware was delivered.

(c) Many applications, especially those relating to non-numeric computing, were out of the range of the hardware.

(d) Little experience had been built up in the development of large, modern software systems.

(e) Computers were not accessible to users who had not had much technical training.

By the late 1960s, internal economic and external military pressures had forced the Soviets to reevaluate their position on computing. By the end of the Eighth Five-Year Plan (1966-1970), computing had become the centerpiece technology of a major campaign to modernize the economy and increase factor productivity. Considerable resources were committed to an effort to upgrade this technology.

The most pressing technical need was for an upward compatible family of general purpose computers. Two Soviet attempts to produce such a family, the Ural-10 series and the ASVT-M models, had both essentially failed by 1970. A third family, the Unified System (Ryad), has done much better. A joint CEMA (Council for Economic Mutual Assistance, the economic counterpart of the Warsaw Pact) effort, Ryad is effectively a reverse engineering* of the IBM S/360 series. Production of the Ryad models began in 1972, and the Soviets and their partners are now working on a new group of machines that closely resemble the IBM S/370 models [3].

The decision to copy the S/360 architecture was made during 1968-69. The East Germans were the major advocate of this course of action. This decision was a reflection of the East-West software gap. None of the CEMA countries had had much experience in developing large, modern software systems nor had they accumulated a large, economically significant collection of applications software. The plan was to appropriate the IBM S/360 operating systems. Then with this base of systems software, they would be in a position to borrow the huge quantities of other systems and applications software that had been developed over the years by IBM and its customers. This plan has been followed with considerable success and

*That is, quantity production of a close functional equivalent has been achieved at reasonable cost.

represents one of the most impressive technology transfers in Soviet history.

Although Ryad, and other Soviet and East European, computer developments are still backward by current Western and Japanese standards, they represent considerable progress compared to what was previously available. They do much to correct the hardware deficiencies noted earlier, and they provide the Soviet Bloc with a respectable and uniform hardware and software base.

But these developments do not in themselves solve the Soviets' computer problems. The newly available technology needs to be effectively integrated into the economy. People have to learn how to make computers do useful things. There is only so much that the Soviets can borrow from IBM and other companies. Ultimately, their success in using computing will depend on how well they can produce, maintain, diffuse and use their own software.

On the surface, it would appear that the Soviets should have an easier time with software technology than with most others. They have a large and distinguished mathematics and theoretical engineering community. Given the compatibility between the Unified System and IBM mainframes, borrowing should be particularly easy. Furthermore, software development avoids two of the worst problems that plague the entire Soviet economy. Given a half-way decent computer installation, and these now exist in some quantity in the USSR, software production does not require a continuing and timely material supply. And given a respectable prototype, the production of quality copies is both cheap and easy. Finally, since the State virtually guarantees employment for all whom it considers at least politically docile, one would think that nobody would view the computer as a threat to his job. In fact, in theory the workers should be falling over themselves in their enthusiasm to find new applications for computers.

Unfortunately for the Soviets, software also preys to an unusual extent on three chronic weaknesses of their economy: customer relations, effective product diffusion and maintenance.

Like most of the other sectors of the economy, software production has to contend with a major behavioral obstacle. Soviet organizations with similar interests tend not to cooperate or interact with each other. Tradition, institutional structure and incentives are such that enterprises try to mind their own business as much as possible. Much of the cooperation that does exist is forced by Party or military demands or by desperate efforts to circumvent supply foul-ups. Other efforts rarely come to much. This has particularly affected software diffusion. Before Ryad, hardware manufacturers did little to produce, upgrade or distribute software. Few models existed in sufficient numbers to make possible a common software base of real economic importance. Repeated attempts to form user groups amounted to little. Soviet security constraints restricted who could participate in sharing software for some models. Enterprises rarely exchanged programs. Contracts with research institutes to produce software products are often frustrating for the customer. The research institute staff may be content with a prototype system that is not well tailored to the customer's needs. Most users have little recourse but to modify and maintain the programs on their own.

Conditions are gradually improving, but changes take time even where they are possible. One promising reform has been the establishment of the corporation-like production associations. These support the creation of relatively large, efficient computer centers that should be able to better serve the needs of the association and its component enterprises. The association may contain a research institute with its own software group. On the surface, at least, an association appears to be a more viable unit for the production and utilization of software, and one that might be able to deal more effectively with other firms. However, seemingly reasonable reforms in the past have actually produced results totally opposite of those that were intended. It is as yet too early to evaluate the impact of this reorganization, either in general or with respect to software development.

In the United States there are a large number of companies that provide professional software services to customers. They range in size from giants like IBM to one-man firms. Some build systems and then convince users to buy them. Others ascertain customer needs, and then arrange to satisfy them. A wide variety of other services are also offered. Basically they are all trying to make a profit by showing their customers how to better utilize computers. To a considerable extent, the software vendors and service bureaus have created a market for themselves through aggressive selling and the competitive, customer oriented, development of general purpose and tailor-made products. The nature of software makes it relatively easy for one firm to make its product better than that of a competitor, who will then figure out a way to upgrade its system to gain a market advantage, etc. There is probably no other sector of the American economy with such a rapid rate of incremental innovation.* The best firms make fortunes, the worst go out of business. Adam Smith would have been overjoyed with this industry.

The Soviets appear to have no real counterpart to these firms for the customer oriented design, development, diffusion and maintenance of software. One enterprise, the Tsentroprogrammisistem Scientific-Production Association in Kalinin, has been publicly identified as a producer of Ryad user software [6]. This organization is under the Ministry of Instrument Construction, Means of Automation and Control Systems (Minpribor). We assume that the Ministry of the Radio Industry, the manufacturer of Ryad in the USSR, has some central software facilities available because of legal responsibilities. Some research institutes, computer factories and local organizations develop and service software, but complaints about their work is common and praise is rare. We know little about what any of these places are doing or how they function. The average Soviet computer user does not seem to have many places it can turn to for help. This should be particularly true of those installations that are not near a major metropolitan area.

The mere fact that we know so little about Soviet software firms is strong evidence that the volume and pace of their activities must be much below that of the American companies, or at least that benefits to users are limited by a lack of readily available information. Most American computer users are not very sophisticated and need to have their hands held by vendors and service companies. There is every reason to believe that most Soviet users are far less sophisticated. It is inconceivable that the USSR has anything comparable to the American software

*Unfortunately, there appears to be no study of the US software industry that would enable us to be more specific.

companies that we do not know about, because then there is no way for the thousands of computer users in the Soviet Union to know about such services either. It is simply not the sort of thing that can be successfully carried on in secret. It must be open and aggressive or it will not reach its customers. It must advertise in some way because most customers are not capable of thinking of useful products on their own.

What is likely is that Soviet installations are pretty much on their own with regard to applications software. The open literature seems to confirm this with articles on how Such-and-Such Production Enterprise built an applications system for itself. There are almost no articles on how some research institute built something like a data base management system that is now being used at scores of installations in a variety of ways. Currently, Soviet installations are building lots of fairly obvious local systems. This pace may actually slow down once these are up and running because there are no effective mechanisms for showing users what they might do next.

Before Ryad the dissemination of software products and services was accomplished through hardware vendors, user groups, informal trades, national and regional program libraries and conferences, and various contractual arrangements. None of this was particularly effective or well organized. For example, the libraries were little more than mail-in depositories that were not properly staffed, indexed or quality controlled. The development of the Unified System was accompanied by a greater appreciation of the limitations of past practices. Ryad hardware would be pitifully underutilized if each user installation were left with an almost empty machine and expected to do all its own programming. This would have defeated the whole purpose of the new system. Improved software products and services would have to be made available to the general community of computer users.

Ryad has brought some real progress in all of these areas, and overall capabilities are improving steadily. However, progress is slow and some important systemic changes may be necessary before it can be greatly accelerated.

The Soviets claim to have "socialized knowledge" and it is thus easier to diffuse scientific and technical information in the USSR than it is in the capitalist countries. "Soviet enterprises are all public organizations, and their technological attainments are open and available to all members of society, with the exception of course of information classified for military or political reasons. The public nature of technological knowledge contrasts with the commercial secrecy that is part of the tradition of private property in capitalist countries. Soviet enterprises are obliged not only to make their attainments available to other enterprises that may wish to employ them but also actively to disseminate to other enterprises knowledge gained from their own innovation experience. The State itself subsidizes and promotes the dissemination of technological knowledge through the massive publication services of the All-Union Institute for Scientific and Technical Information [VNITI]" [1]. This sounds better in theory than it works in practice. While services like those provided by VNITI are unquestionably useful, they do not compare to the much broader range of diffusion services available in the US. Capitalistic commercial secrecy is overstated; very little remains secret for very long. The Soviets have no real counterpart for the volume and level of Western marketing activity. By comparison lists of abstracts of products that have not been properly quality controlled for market conditions, that have no real guarantees or back-up service, etc. cannot be expected to be as effective a vehicle for diffusion. The Soviet incentive structure not only does not encourage dissemination of innovation particularly well, but it also often promotes the concealment of an enterprise's true capabilities from its superiors. Few feel nobly obliged to make their attainments available to anyone else.

The vertical structuring of the Soviet ministerial system works against software development and diffusion. Responsibility is primarily to one's ministry and communication is up and down ministerial lines. It is much easier to draw up economic plans for this kind of structure than it is for those with uncontrolled horizontal communication. Furthermore, each ministry appears determined to retain full control of the computing facilities used by its enterprises. In the West, software diffusion is a decidedly horizontal activity. Data processing and computing personnel and management talk to each other directly across company and industry lines. This communication is facilitated by very active professional organizations. Such arrangements do not exist to anywhere near the same extent in the USSR.

It is, of course, not only the ministerial system that mitigates against the really effective encouragement of direct producer-customer horizontal economic activity. Often the various layers of local Communist Party organizations perform the role of facilitating horizontal exchanges. The Party needs visible activities that justify its existence and authority, and this is one of the most important. No serious erosion of this prerogative is possible. However, it is much easier for a local Party secretary to get a carload of lumber shipped than it is for him to expedite the delivery of a special purpose real-time software system. He can take the lumber away from a lower priority enterprise, but what can he do to get the bugs out of the software? He can throw extra people on the job, but that will probably only make matters worse. Software projects tend to react badly to the "Mongolian horde" approach often favored by the Soviets. The detailed enterprise level software transactions cannot be managed by politicians.

This problem affects the diffusion of technical R & D to production enterprises in general. Software is an extreme case because it is so difficult to manage under any circumstances. One mechanism that has evolved to facilitate technical work is the emergence of very large enterprises and research institutes that are capable of handling most of their own needs in-house. Thus one finds many enterprises who own and operate computing facilities entirely on their own. This is basically a defensive reaction that improves local viability in the hostile ministry/Party environment. Globally, the wide distribution, limited use, and hoarding of scarce resources, particularly personnel, in bloated organizations is counterproductive. The Party and government do recognize this and have shown themselves prepared to give up some control to obtain increased efficiency in innovation. Most of these changes have related to highly technical R & D matters over which they have had little effective control anyway. Changes include the already discussed corporation-like associations and R & D contract work, and also reforms in innovation incentives and prices for new products. This represents progress and will help the development and diffusion of software. It still falls far short of the systemic advantages enjoyed by the US software industry.

In summary, the Soviets have made considerable progress in removing limitations due to hardware availability, some progress as a result of changes in priorities, and as yet relatively little progress in overcoming an assortment of complex systemic problems that affect the development of software. Consequently, the USSR will have to continue to borrow from foreign software technology, and it is now better equipped and motivated to do so.

REFERENCES

[1] Berliner, Joseph S., The Innovation Decision in Soviet Industry, MIT Press, Cambridge, Mass., 1976.

[2] Boehm, Barry W., "Software Engineering: R & D Trends and Defense Needs", pp. 1.1-1.43 in Wegner, Peter (ed.), Proc. Conf. on Research Directions in Software Technology, MIT Press, Cambridge, Mass., 1978. The author's copy is a 1977 preprint from the conference.

[3] Davis, N. C. and S. E. Goodman, "The Soviet Bloc's Unified System of Computers", ACM Computing Surveys, Vol. 10, No. 2, June, 1978, 93-122.

[4] Goodman, S. E., "Software Development in the USSR", to appear as a Department of Defense report under the auspices of the U. S. Army Foreign Science and Technology Center, Charlottesville, Va., 1978.

[5] Goodman, S. E., "Software Technology Transfer to the USSR", to appear as a Department of Defense report under the auspices of the U. S. Army Foreign Science and Technology Center, Charlottesville, Va., 1978.

[6] Myasnikov, V. A., "Results and Priority Tasks in the Field of Automation of Control Processes in the National Economy of the USSR", Upravlyayushchiye Sistemy i Mashiny, Kiev, No. 1, Jan.-Feb., 1977, 3-6.

References [1,3,4] contain large bibliographies.

A MARTINGALE THEORY OF RANDOM FIELDS

E. Wong
University of California, Berkeley
Department of Electrical Engineering and Computer Science
Berkeley, California 94720

ABSTRACT. With a likelihood ratio problem as a focal point, recent development in the theory of martingales and stochastic calculus for two dimensional random fields is reviewed.

I. INTRODUCTION.

By a random field we shall mean a stochastic process $\{X_t, t \in T\}$ with a multidimensional parameter set T. In this paper we shall outline a theory of martingales and stochastic integration for random fields with a two-dimensional parameter, and in the process review the recent developments in this area.

For the one dimensional case, a martingale is defined as follows: Let $\{\mathcal{F}_t, 0 \leq t \leq T\}$ be an increasing family of σ-fields. A stochastic process M_t is said to be an $\mathcal{F}_t$-martingale if

$$(1.1) \qquad E(M_{t+s} | \mathcal{F}_t) = M_t \quad \text{a.s. for all } s \geq 0$$

or equivalently M_t is $\mathcal{F}_t$-measurable for every t and

$$(1.2) \qquad E(M_{t+s} - M_t | \mathcal{F}_t) = 0 \quad \text{a.s. for all } s \geq 0.$$

Observe that since $M_t = E(M_T | \mathcal{F}_t)$, martingale theory is essentially a theory of the dynamics of $\{\mathcal{F}_t\}$, which in applications has the interpretation of being the information dynamics. Furthermore, (1.2) is an expression of a quasi-independence for forward increments, and this fact is responsible for a martingale calculus with both simplicity and power.

The motivation for the two dimensional case is similar, though not at first glance. For example, since the parameter is now no longer identified as time, it is not clear as to where the "dynamics" comes from and why it is important. To make the motivation explicit, consider a specific information processing problem involving two-dimensional data described as follows: Suppose that a random field η_t is observed for $t = (t_1, t_2)$ on a rectangle $T = [0,T_1] \times [0,T_2]$. We want to decide between the two hypotheses:

H_0 : η_t is a Guassian white noise

H_1 : $\eta_t = S_t + \zeta_t$ where S_t is a random signal and ζ_t is a Guassian white noise

From the Neyman-Pearson lemma we know that the statistic that we need to compute is the likelihood ratio

$$(1.3) \qquad L = E_0 \left(\frac{d\mathcal{P}}{d\mathcal{P}_0} | \eta_t, t \in T\right)$$

where $\mathcal{P}$ and $\mathcal{P}_0$ are the probability measures under H and H_0 respectively and

$\frac{d\mathcal{P}}{d\mathcal{P}_0}$ is the Radon-Nikodym derivative of $\mathcal{P}$ w.r.t. $\mathcal{P}_0$. The problem has now been reduced to one of finding an explicit expression of L as a functional of $\{\eta_t, \ t \in T\}$, and finding an effective means for computing it.

If S_t is a deterministic signal then it is not difficult to show that L is given by

$$L = \exp\left\{\int_T S_t \, \eta_t \, dt - \frac{1}{2}\int_T S_t^2 \, dt\right\}$$

where the white noise integral is easily defined as a Wiener integral. If S_t is a Gaussian process, then an expression for L can be derived using either the Karhuven-Loeve expansion or reproducing kernel Hilbert space techniques. The resulting formula, however, does not lend itself to efficient computation. In the general case where S_t is random and non-Gaussian no expression for L was known until one was derived using martingale techniques. [5]

II. THE ONE-DIMENSIONAL CASE . To see why martingale theory is useful for the likelihood ratio problem, consider the situation in one dimension. We can use the same problem description and retain the same notation. The only difference is that t and T are now one-dimensional. Define

$$L_t = E_0 \left(\frac{d\mathcal{P}}{d\mathcal{P}_0} \middle| \mathcal{F}_{\eta t}\right)$$

where $\mathcal{F}_{\eta t} = \sigma\{\eta_s, \ 0 \le s \le t\}$. It is clear that the likelihood ratio L is just L_T, and that L_t is a $(\mathcal{P}_0, \mathcal{F}_{\eta t})$ martingale. Hence, the likelihood ratio has been embedded in a martingale.

The results in the one-dimensional case can be summarized as follows: Let Y_t be a Brownian motion process defined by $Y_t = \int_0^t \eta_\tau \, d\tau$. Then L_t satisfies the integral equation.

$$L_t = 1 + \int_0^t L_\tau \hat{S}_\tau \, dY_\tau \tag{2.1}$$

which in turn can be solved to yield the explicit formula

$$L_t = \exp\left\{\int_0^t \hat{S}_\tau \, dY_\tau - \frac{1}{2}\int_0^t \hat{S}_\tau^2 \, d\tau\right\} \tag{2.2}$$

In both of these equations the integral involving Y is defined as an It$_0$ integral, and $\hat{S}$ is defined by

$$\hat{S}_t = E(S_t | \mathcal{F}_{\eta t}) \tag{2.3}$$

That is, $\hat{S}_t$ is just the causal estimator of the signal given the observation.

The equations (2.1) and (2.2) are interesting from many points of view, some of which can be summarized as follows:

(a) The likelihood ratio L is given by

$$L = \exp\left\{\int_0^T \hat{S}_\tau dY_\tau - \frac{1}{2}\int_0^T \hat{S}_\tau^2 \, d\tau\right\} \tag{2.4}$$

where T appears only as the end point in the integrals.

(b) The problem of computing L has been reduced to the filtering problem of computing $\hat{S}_t$.

(c) Since (2.1) implies that

$$dL_t = L_t \hat{S}_t \, dY_t$$

L can be recursively computed whenever $\hat{S}_t$ can be so computed.

(d) Even though L itself involves no dynamics, the expression (2.4) for L was obtained only because the problem was embedded in a dynamical formulation.

III. TWO-PARAMETER MARTINGALES AND STOCHASTIC CALCULUS. Our obejctive is to obtain likelihood ratio formulas similar to (2.1) and (2.2) for the two dimensional case. The experience in the one-dimensional case suggests that what we need is a theory of martingales and stochastic calculus which takes full advantage of the white noise sturcture.

Let ζ_t be a white Gaussian noise with a two-dimensional parameter, so that

(3.1) $$E\zeta_t \zeta_s = \delta(t-s) = \delta(t_1 - s_1)\,\delta(t_2 - s_2)$$

Let $R_+^2 = \{t : 0 \leq t, < \infty, \ 0 \leq t, < \infty\}$ denote the positive quadrant of the plane. For a Borel set A in R_+^2 consider

(3.2) $$W(A) = \int_A \zeta_t \, dt$$

The set-parameterized process W is Gaussian with zero mean and

(3.3) $$EW(A)\,W(B) = \text{Area } (A \cap B)$$

We shall call W the standard Wiener process with a two dimensional parameter.

For a pair of points t and s in R_+^2 define a partial ordering by

$$t \succ s \iff t_1 \geq s_1 \text{ and } t_2 \geq s_2$$

A family of σ-fields $\{\mathfrak{F}_t, \ t \in R_+^2\}$ is said to be increasing if

$$t \succ s \Longrightarrow \mathfrak{F}_t \supset \mathfrak{F}_s$$

Given an increasing family of σ-fields $\{\mathfrak{F}_t\}$, a process $\{M_t. \ t \in R_+^2\}$ is said to be an $\mathfrak{F}_t$-martingale if

(3.4) $$t \succ s \Longrightarrow E(M_t | \mathfrak{F}_s) = M_s \quad \text{a.s.}$$

This definition generalizes (1.1). M is said to be an adapted weak martingale if M_t is $\mathcal{F}_t$-measurable for each t and

(3.5) $$t \succ s \Longrightarrow E[M_t - M_{t\otimes s} - M_{s\otimes t} + M_s | \mathcal{F}_s] = 0 \quad \text{a.s.}$$

where $t\otimes s = (t_1,s_2)$. This definition generalizes (2.2). A martingale is always an adapted weak martingale but the converse need not be true. [1]

Let W be a standard Wiener process and let $\{\mathcal{F}_t, t \in R_+^2\}$ be an increasing family of σ-fields. Let A_t denote the rectangle $\{s : s \in R_+^2 \text{ and } s \prec t\}$, and consider the process $W_t = W(A_t)$. W_t is an $\mathcal{F}_t$-martingale if W_t is $\mathcal{F}_t$-measurable for every t and $W(\Delta)$ is $\mathcal{F}_t$ independent whenever $\Delta \cap A_t = \phi$. We shall denote this situation by saying that $\{W_t, \mathcal{F}_t, t \in R_+^2\}$ is a Wiener process.

In one dimension the stochastic calculus associated with a Gaussian white noise is based on the Ito integral. A rather strightforward generalization of the Ito integral for the two dimensional case can be used to define integrals of the form

$$\int_A \psi_t \, W(dt)$$

where ψ is a measurable process adapted to $\{\mathcal{F}_t\}$ and square-integrable with respect to $d\mathcal{P} \times dt$. Roughly speaking, the integral is the limit of "forward difference" approximations $\sum_\nu \psi(t_\nu)\, W(\Delta t_\nu)$ where Δt_ν is an incremental area with points $s \succ t$.

Analogy with the one dimensional case suggests the following questions:

(a) Martingale

If we define a process M_t by

(3.6) $$M_t = M_0 + \int_{A_t} \psi_s W(ds)$$

then is M a martingale? The answer is yes.

(b) Completeness

Let $\{\mathcal{F}_t\}$ be generated by the Wiener process W, i.e., $\mathcal{F}_t = \sigma(W_s, s < t)$, and let M be a square-integrable $\mathcal{F}_t$ martingale. Is M_t necessarily of the form (3.6)? The answer to this question is no. In general, we need stochastic integrals of a second kind

(3.7) $$\int_{A_t \times A_t} \psi_{s,s'} \, W(ds)\, W(ds')$$

whose definition and details of the completeness result can be found in [2].

(c) Closure

Let X_t be a process of the form

(3.8) $$X_t = X_0 + \int_{A_t} \theta_s \, ds + \int_{A_t} \psi_s \, W(ds) + \int_{A_t \times A_t} \psi_{s,s'} \, W(ds)\, W(ds')$$

and let f be a smooth function, is $f(X_t)$ again of the form (3.8)? Once more, the answer is no. To achieve closure, we need mixed integrals

$$\int_{A_t \times A_t} \alpha_{s,s'}\, W(ds)\, ds' \quad \text{and} \quad \int_{A_t \times A_t} \beta_{s,s'}\, ds\, W(ds')$$

which are defined in [4]. The closure result is in the form of a differentiation formula and is given in [6].

IV. LIKELIHOOD RATIO FORMULAS. Let us return to the problem of determining the likelihood ratio L as defined by (1.3). Let

$$Y_t = \int_{A_t} \eta_s\, ds \tag{4.1}$$

Then under $\mathcal{P}_0$ Y_t is a Wiener process. Let $\mathcal{F}_{yt} = \sigma(Y_s, s < t)$ and define

$$L_t = E_0\left\{\frac{d\mathcal{P}}{d\mathcal{P}_0} \middle| \mathcal{F}_{yt}\right\} \tag{4.2}$$

The, L_t is a $(\mathcal{P}_0, \mathcal{F}_{yt})$ martingale and the completeness result yields the formula. [3]

$$\begin{aligned} L_t = 1 &+ \int_{A_t} \hat{S}(\tau|\tau)\, L_\tau\, Y(d\tau) \\ &+ \int_{A_t \times A_t} R(\tau,\tau'|\tau v\tau')\, L_{\tau v\tau'}\, Y(d\tau)\, Y(d\tau') \end{aligned} \tag{4.3}$$

where $\hat{S}$, R and $\tau v\tau'$ are defined as follows:

$$\hat{S}(\tau|t) = E(S_\tau|\mathcal{F}_{yt}) \tag{4.4}$$

$$\begin{aligned} R(\tau,\tau'|t) &= E[S_\tau S_{\tau'}|\mathcal{F}_{yt}] \\ &= \rho(\tau,\tau'|t) + \hat{S}(\tau|t)\, \hat{S}(\tau'|t) \end{aligned} \tag{4.5}$$

$$\tau v\tau' = (\max(\tau_1, \tau_1'), \max(\tau_2, \tau_2')) \tag{4.6}$$

Equation (4.3) is an integral equation for L_t which generalizes (2.1).

In [5] the integral equation (4.3) was solved to yield the following explicit formula for L_t:

$$\begin{aligned} L_t = \exp\Bigg\{ &\int_{A_t} \hat{S}(\tau|\tau)\, Y(d\tau) - \frac{1}{2} \int_{A_t} \hat{S}^2(\tau|\tau)\, d\tau \\ &+ \int_{A_t \times A_t} \rho(\tau,\tau'|\tau v\tau')\, [dY_\tau - \hat{S}(\tau|\tau v\tau')d\tau][dY_{\tau'} - \hat{S}(\tau'|\tau v\tau')] \end{aligned} \tag{4.7}$$

$$- \frac{1}{2} \int_{A_t \times A_t} \rho^2(\tau,\tau'|\tau\vee\tau')\, d\tau\, d\tau' \Big\}$$

Equation (4.7) is a full generalization of (2.2), and our objective is attained.

Equations (4.3) and (4.7) show that the problem of computing the likelihood ratio can be reduced to that of computing $\hat{S}$ and ρ. Once again the hypothesis testing problem is reduced to a filtering problem. If S is Gaussian, then the covariance ρ is necessarily a deterministic function, and if $\hat{S}$ can be computed recursively, then so can L. In [7] such a recursive procedure for computing $\hat{S}$ is given.

References

1. Cairoli, R. and Walsh, J.B. (1975). Stochastic integrals in the plane. Acta Mathematica 134, 111-183.

2. Wong, E. and Zakai, M. (1974). Martingales and stochastic integrals for processes with a multidimensional parameter. Zeit. Wahrscheinlichkeitstheorie 29, 109-122.

3. Wong, E. (1974). A likelihood ratio formula for two-dimensional random fields. IEEE Trans. Information Theory 20, 418-422.

4. Wong, E. and Zakai, M. (1976). Weak martingales and stochastic integrals in the plane. Ann. Probab. 4, 570-586.

5. Wong, E. and Zakai, M. (1977). Likelihood ratios and transformation of probability associated with two-parameter Wiener processes. Zeit. Wahrscheinlichkeitstheorie 40, 283-308.

6. Wong, E. and Zakai, M. (1978). Differentiation formulas for stochastic integrals in the plane. Stochastic Processes and Their Applications 6, 399-349.

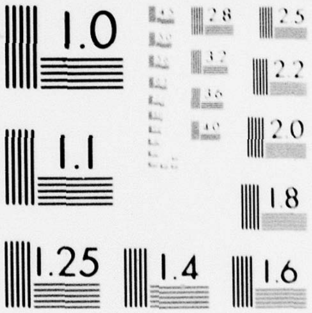
7. Wong, E. (1978). Recursive causal linear filtering for two-dimensional random fields. IEEE Trans. Information Theory 24, 50-59.

AD-A064 433 ARMY RESEARCH OFFICE RESEARCH TRIANGLE PARK N C F/G 12/1
TRANSACTIONS OF THE CONFERENCE OF ARMY MATHEMATICIANS(24TH) HEL--ETC(U)
JAN 79

UNCLASSIFIED ARO-79-1 NL

6 OF 6
AD
A064433

END
DATE
FILMED
4-79
DDC



MICROCOPY RESOLUTION TEST CHART

OPTIMIZATION OF THE MEMORY CAPACITY OF A STORE AND FORWARD RELAY*

W. Pressman and J. Benson**
Communications/Automatic Data Processing Laboratory
US Army Electronics Command
Fort Monmouth, New Jersey 07703

ABSTRACT

It is shown that an array transmitting fixed format messages at random over several channels to a single data relay can be modeled as an appropriate queue. Bounds for the system characteristics are calculated for a wide variety of traffic conditions and compared with other models. In particular, the percentage of messages lost is expressed as a function of the finite memory length and the ratio of the mean arrival and service times.

*This research was supported under the In-House Laboratory Independent Research program.

**The authors of this paper presented it at the 23rd Conference of Army Mathematicians.

1. Introduction

The problem to be considered is as follows: A data relay under design is to be used in a proposed field artillery acoustic location system. In this system acoustical sensors are distributed in an array at known positions beyond the forward edge of the battle area (FEBA). The time of arrival and possibly other signal characteristics of an acoustic signal generated by an explosion are transmitted in a fixed format message from the sensors to a central processing unit on the friendly side of the FEBA, which then evaluates the location of the enemy artillery and time of firing. At extended ranges a relay is necessary. Great emphasis is placed on minimizing the weight and volume of the relay as it must be implanted not only by land or aircraft but also possibly delivered by artillery shell. In addition, the smaller the size of the relay the lower the chances of detecting it.

Several (say N) transmission channels from the sensors to the relays are required in order to minimize radio-frequency overlap and consequent loss of the message at the receiving antennas of the relay. In passing, we mention that a similar problem occurs in the possible overlapping of acoustic signals received at a single sensor and also in the RF message overlap at the CPU. These two problems can be treated analogously and the percentage loss of messages calculated, although we do not do that here.

At the relay, the N channels are electronically scanned and their contents merged into a single memory queue of length K. Effectively, this procedure can be considered to be instantaneous relative to the time scale we are concerned with. The messages are then selected in accordance with some queueing discipline, say FIFO*, time-tagged, error-coded and retransmitted as another fixed format message. Thus, the service time is deterministic. Our basic concern is to determine the number (or percentage) of received messages which are lost before retransmission because of the finite storage capacity of the relay.

2. Formulation of Problem

We now formulate the problem more carefully. The nomenclature we will use is standard in the literature on queueing theory.

First, we postulate that the number of messages arriving per unit time at each channel follows a Poisson distribution with mean arrival rate, b_i, per unit time. (This is equivalent to stating that the inter arrival times are distributed exponentially with average time between arrivals equal to $1/b_i$ (see [1] pages 23-29 for a proof). Further, consider the N independent channel sources of messages where for channel i the mean number of messages per unit time is b_i. It can be proved that the simple queue formed by merging the input from each of the N

* First in, first out. See Section 4 after equation (7) for discussion of queueing discipline.

sources is also Poisson with parameter $b = b_1 + b_2 + \ldots + b_n$. See [2] for a proof.

Thus, we can now formulate our mathematical model as a single queue (1) with exponential arrivals (M) at the memory bank having a finite storage capacity (K) and a fixed deterministic service time (D), the retransmission message length.

In standard notation, it is an M/D/1/K queueing model. We comment on this notation. In general, the entries in the first and second spaces designate the arrival and service-time distributions respectively. The entry in the third space gives the number of parallel service channels, and the last entry signifies the maximum queue capacity.

For later use, we give the Poisson and exponential frequency distributions, their means and variance in Figure 1.

3. The General Approach

We have been unable to find a treatment of the specific model M/D/1/K in the literature and leave its analytic investigation to a future time. Instead, we will study the statistics of three other mathematical models: M/M/1/oo, M/M/1/K, and M/D/1/oo. We show the relationship among these models and M/D/1/K as a function of queue capacity and service frequency distribution in Figure 2. The reason we study these models is that their important statistical properties have previously been deduced and they are "neighboring" models of M/D/1/K. In particular, we shall focus on the queue M/M/1/K because it is a general rule of queueing theory that when a deterministic parameter is replaced by a probabilistic parameter things "get worse". See Kleinrock[2], pp. 188, 191 for a general discussion and some specific examples. In our case, the deterministic serving rate m with variance zero of the queue M/D/1/K is replaced by the exponential serving rate with mean also equal to m, but with variance equal to that of the model M/M/1/K. See Figure 3.

Therefore, the results we obtain from the M/M/1/K model are conservative and will place an upper bound, for example, on the percentages of messages lost for specified values of the data storage capacity. For utility and in order to determine the sensitivity of our model, we evaluate the model statistics for a wide range of the parameters, r and K. Here, K has been defined previously and r is the traffic intensity

$$r = b/m, \qquad (1)$$

where b is the mean arrival rate and m is the mean service rate.

In section 4, we perform the analysis and summarize the results. In section 5, we apply our results and comment on the need for field measurements to validate the model and estimate the parameters. We conclude with a remark on the usefulness of powerful, modern simulation techniques now available.

4. Analysis

We list in Figures 4-6 the relevant statistics of the three queue models M/M/1/oo, M/M/1/K, and M/D/1/oo. The equations have been extracted from references [1,2], which also contain the derivations of these formulas. Some comment on the notation used is necessary. We define p_n as the probability of having n messages in the system (queue plus server), rather than the queue alone, as is done in some texts. The important queue parameter r - the traffic intensity - is as defined in equation (1) of section 3. The number of messages in storage fluctuates of course. L_q is the average number in the queue, L the average number in the system, and S_L^2 is the variance of the number in the system from its mean value L.

Examination of Figures 4-6 provides some useful insights. It is surprising, perhaps, how simple the statistics are for the infinite single-server queue with exponential arrivals and service (Figure 4). It is as though the variance of the server rate absorbed much of the variance of the arrival rate so that all statistics can be expressed conveniently in terms of the traffic intensity. Even when the queue length is restricted to be finite (Figure 6), the equations are in closed form and although more complicated are still relatively simple. However, when we consider the infinite single-server queue with exponential arrivals but deterministic-fixed serving rate (Figure 5), M/D/1/oo, the "mismatch" causes difficulty. Note that the formula for p_n becomes ever larger and unwieldy as n increases. Thus, although there may exist formulas for p_n in the corresponding model of finite queue length, M/D/1/K, they are likely to be so complicated that approximation methods would probably be necessary in this eventuality.

In section 3, we stated that introducing randomness (variance) makes things (queue statistics) worse. Here we clarify this statement. Pollaczek and Khintchine , (PK) have proved the following theorem [1], pp. 225-226):

For an infinite single-server queue with exponential arrival distribution (M) and general service distribution (G), M/G/1/oo, the average number in the system (L) is

$$L = r + ((r^2 + b^2 s_v^2)/2(1-r)) \qquad (2)$$

Here L, r, and b are as previously defined. The quantity s_v^2 is the variance of the general frequency distribution characterizing the service. If we set G equal to D (i.e. a fixed-deterministic service rate) then s_v is equal to zero and (2) becomes

$$L = r + (r^2 / 2(1-r)) \qquad (3)$$

Combining equations (1) and (2) of Figure 5 gives exactly (3) above. If we let G equal M (i.e. an exponential service rate) then from (6) of Figure 1

$$s_v^2 = 1/m^2 \qquad (4)$$

Inserting (4) in (2) and using (1) gives exactly equation (3) of Figure 4:

Thus, we have verified the validity of the (PK) general formula for the special cases M/D/1/oo and M/M/1/oo. In general, equation (2) states that adding variance to the serving process increases the average number in the system by $b^2 s_v^2/2(1-r)$. From (2) and (3) of Figures 4 and (1) and (2) of Figure 5 we also see directly that

$$L_q\ (M/D/1/\infty) = \tfrac{1}{2} L_q\ (M/M/1/\infty),$$

$$L\ (M/D/1/\infty) = (1 - \frac{r}{2})\ L\ (M/M/1/\infty),$$

$$o < r < 1$$

If we now consider M/D/1/K and M/M/1/K, the two analogous models with finite system capacity K, then we are persuaded by the prior discussion that it is more likely that the Kth slot of the latter model will be occupied than the Kth slot of the first, that is

$$p_k\ (M/M/1/K) \geq p_k\ (M/D/1/K)$$

Suppressing the common queue attributes, we abbreviate the above.

$$p_k\ (M) \geq p_k\ (D) \qquad (5)$$

We are particularly interested in p_k for the following reason. Our major interest is to determine the average number on percentage of messages lost as a function of the traffic intensity, r, and the maximum system capacity K. To find this, we must determine the probability of an arriving message finding the queue filled. When we multiply this by the average arrival rate, b, we obtain bp_k, the expected number of messages which arrive at a filled queue and are not stored. Finally, if we divide this last quantity by the average arrival rate b, we obtain the fraction of messages, p_k, lost on the average (LS). For the model M/D/1/K, we write

$$LS\ (D) = p_k\ (D). \qquad (6)$$

Then from (5)

$$LS\ (D) \leq p_k\ (M). \qquad (7)$$

This is our key equation. It should be noted that although we have specified a FIFO queue disciple in section 1, it is not necessary to do so for our limited purposes. If the queue is filled and a new message arrives it is immaterial to us if this message is lost or another message is displaced from the queue. An alternate queueing discipline says FILO - first in, last out - would yield the same p_k but different waiting time statistics. We are not concerned with this latter problem here.

From equations (1) and (2) of Figure 6 we can compute $p_k(M)$ as function of r and k. We have done this for $0 \leq r \leq 1$ and $k = 1,2,\cdots,10,15,20,40$. Our results are shown in graphical form in Figure 7. In addition to numerical values the graph serves to determine the sensitivity of any one variable to variations in the other two. For more precise calculations, the formula

$$p'_k = (1-r)\, r^k \,/\, (1-r^{\,k+1}) \qquad (8)$$

may be used.

5. <u>Implementation, Final Comments</u>

It is possible to use equation (7), (8) and Figure 7 in a variety of ways for design, planning or operational purposes. For example, we give one design scenario. Suppose, that as a result of field manuevers and/or war-gaming, it is determined that on the average 20 message/sec will arrive at the relay queue, and the restrictions on the contents of the fixed format retransmission message permit 36 messages/sec to be sent. Further, it is required that no more than 5% of the messages be lost due to memory overflow at the relay, otherwise the total system performance will be degraded. What should the minimal data storage capacity (k-1) at the relay be? From the above

$$r = b/m = 20/36 = .555,$$

$$p_k = .05.$$

Then from Figure 7, $k = 4$.
Thus, 3 is a conservative estimate of the number of memory slots for the specified r and p'_k.

We close with two final comments. First: it is only by careful field test and war-gaming based on appropriate scenarios that the M/D/1/K model we have postulated can be validated and the value of b, the mean Poisson arrival rate determined. Second: regardless of the type of frequency distribution of arriving messages which is ultimately determined to be valid, modern simulation techniques on high speed computers are now available to calculate the requisite queue statistics. Shannon (3) in a recent text includes chapters on event-oriented discrete simulation techniques (appropriate for queueing problems) and design of computer experiments. Pritsker (4) has produced a Fortran-based simulation language, GASP <u>IV</u>, quite suitable for queueing models. Here we have used an analytic approach.

Poisson: $P_n = b^n \exp(-b)/n!$. (1)

The mean $\bar{n} = b$, and (2)

The variance $s^2 = b$ (3)

Here F_n is the probability of n arrivals in a time interval of unit length, n being a non-negative integer. The distribution is discrete.

Exponential:

$f(t) = m \exp(-mt)$ (4)

The mean $\bar{t} = 1/m$ and (5)

the variance $s^2 = 1/m^2$ (6)

Here $f(t)$ is the probability of t being the time between successive arrivals. The distribution is continuous.

Figure 1. Poisson and Exponential Distributions

SERVICE DISTRIBUTION	QUEUE CAPACITY	
	Infinite	Finite
Exponential	M / M / 1 / oo	M / M / 1 / K
Deterministic	M / D / 1 / oo	M / D / 1 / K

FIGURE 2. The Four Queue Models

	M / M / 1 / K	M / D / 1 / K
ARRIVAL DISTRIBUTION	Both models are exponential - mean and variance is the same for both models.	
SERVICE DISTRIBUTION	EXPONENTIAL	DETERMINISTIC
Mean	m	m
Variance	m	o

FIGURE 3. Neighboring Queue Models

Traffic Intensity:

$r = a/m$, where a and m are the mean arrival and service rates respectively.

(1) $p_n = (1-r)r^n \quad , \quad n = 0,1,2\ldots$

(2) $L_q = r^2/(1-r)$

(3) $L = r/(1-r)$

(4) $S_L^2 = r/(1-r^2)$

FIGURE 4. Statistics of Queue M/M/1/oo

$$(1) \quad L_q = r^2/2(1-r)$$

$$(2) \quad L = r + L_q$$

$$(3) \quad P_o = 1 - r \ , \ P_1 = (1 - r)(e^r - 1)$$

$$(4) \quad P_2 = (1 - r)(e^{2r} - e^r(1 + r))$$

$$(5) \quad P_3 = (1 - r) \frac{e^r}{2} (r^2 + 2^r) + e^{2r}(1 - 2r) + e^{3r}$$

$$(6)\text{-} \quad P_n = (1 - r)\{ \sum_{i=1}^{n} (-1)^{n-i} e^{ir}[\frac{(ir)^{n-1}}{(n - i)!} + \frac{(ir)^{n-i-1}}{(n - i - 1)!}] \} \quad *$$

* Exclude i = n in second term.

Figure 5. Statistics of Queue M/D/1/∞

(1) $P_n = \dfrac{(1-r)r^n}{(1-r^{k+1})}$ $(r \neq 1)$

(2) $P_n = \dfrac{1}{k+1}$ $(r = 1)$

$n=0,1,\ldots,K$

(3) $L = r\left[\dfrac{1-(k+1)r^{k} + kr^{k+1}}{(1-r^{k+1})\,(1-r)}\right]$

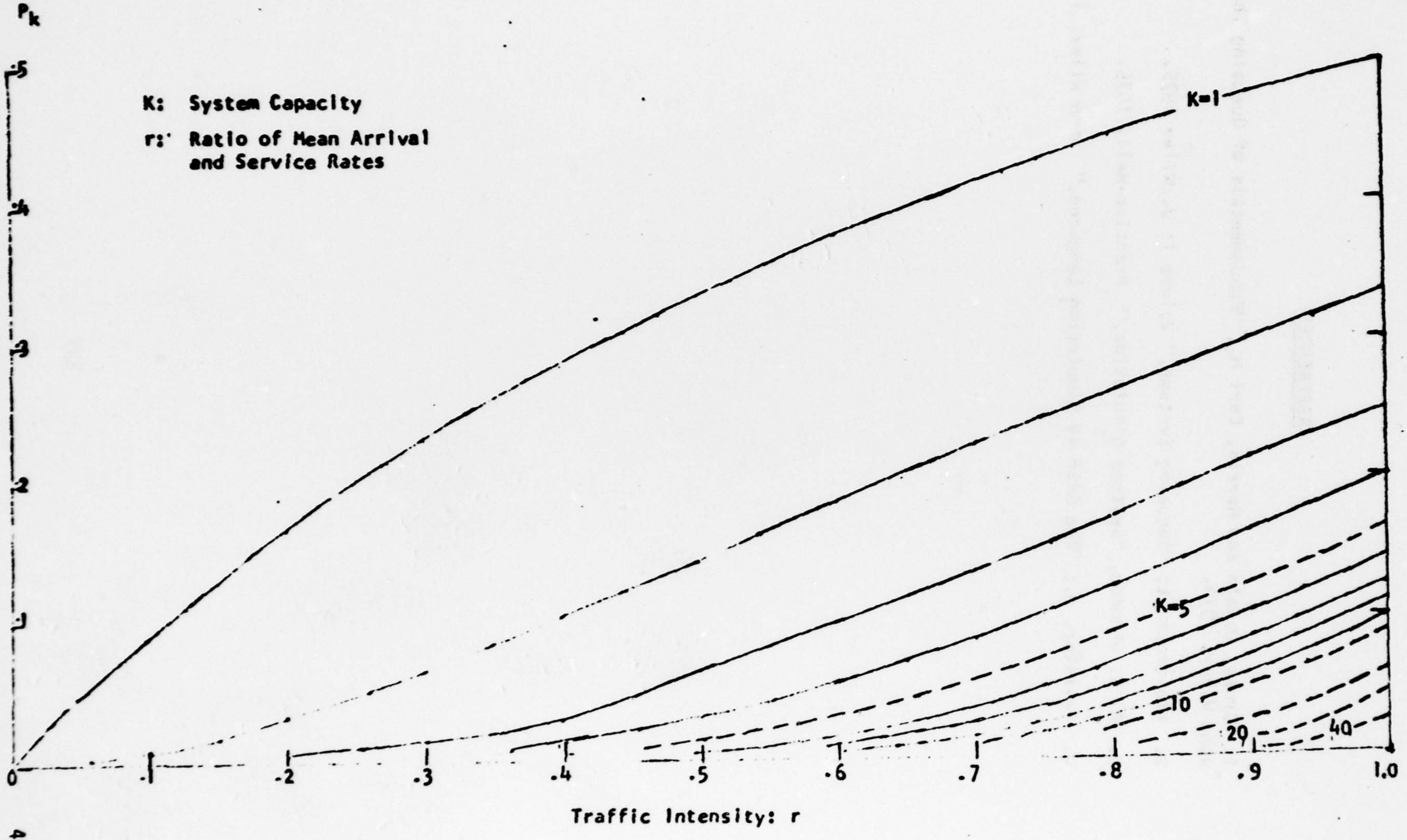
$L_q = L - \dfrac{r\,(1-r^{k})}{1-r^{k+1}}$

(4) Average Fraction of
<u>Messages</u> <u>Lost</u> $= P_k$

Note: k is system capacity: queue capacity plus one (the server)

Figure 6. Statistics of Queue M/M/1/K

FIGURE 7. FRACTION OF INPUT MESSAGES TO QUEUE M/M/1/K WHICH IS LOST (P_k)



REFERENCES

1. Gross, Donald and Harris, Carl M.; "Fundamentals of Queueing Theory;" John Wiley, 1974.

2. Kleinrock, L; "Queueing Systems," Volume 1; J. Wiley 1975.

3. R. E. Shannon, "Systems Simulation," Prentice-Hall, 1975.

4. Pritsker, A.; "The GASP IV Simulation Language," John Wiley, 1974.

LIST OF ATTENDEES

24th Conference of Army Mathematicians
31 May - 2 June 1978
Charlottesville, Virginia

BERGER, Marc A.	U of Wisc (MRC)
BLOOMFIELD, R. S.	U of VA
BOWDEN, Charles M.	Redstone Arsenal
BRAVY, Steve	Concepts Analysis Agency
BROCKETT, Roger W.	Harvard U
CAMPBELL, Thomas W.	FSTC
CHANDRA, Jagdish	ARO
CHARTRES, Bruce A.	U of VA
CHOW, Pao L	Wayne State U
COFFEE, Terence	BRL
COHEN, Edgar A., Jr.	NSWC/WOL
COLANTRONI, G	U of VA
COLEMAN, Norman P., Jr.	ARRADCOM
COLLIER, Alan G.	FSTC
COUNCIL, F. E., Jr.	MISO
CROISANT, William J.	ACERL
DAVIS, Julian L.	ARRADCOM
De BOOR, Carl	U of Wisc (MRC)
Di Perna, Ronald J.	U of Wisc (MRC)
DEWISPELLING, Aaron	U of VA
DOUGLAS, A. J.	Dept of National Defense, Canada
ECKER, J. G.	Rennselear Polytech
FIX, Grace A.	TARADCOM
GALBRAITH, A. S.	ARO (Retired)
GAMBINO, Lawrence A.	CSL
GOLDSTEIN, Marvin	NUSC
GOLOMB, Michael	U of Wisc (MRC)
GOODMAN, S. E.	Princeton U
GRASSI, R. (LTC)	FSTC
GREEN, Robert A.	Defense Communication Agency
GREVILLE, Thomas N. E.	U of Wisc (MRC)
HARRINGTON, David P.	U of VA
HEIMERL, Joseph M.	BRL
HIGHFILL, J. H.	U of VA
JOHNSON, Ralph	CAA
KALMAN, R. E.	U of Florida
KEEVER, David	U of VA
KOLEYNI, Ghassem	U of VA
KOSIEWICZ, John J.	FSTC
KOTIN, Leon	CORADCOM
KRING, Jonathan F.	NASA

LAUNER, Robert L.	ARO
LeVAN, M. Douglas	U of VA
LIN, Y. K.	U of ILL
MAHLER, Carl P. B.	U of VA
MANN, James E., Jr.	U of VA
McALPINE, G. A.	U of VA
McSHANE, E. J.	U of VA
MEALS, L. Kenton	David W. Taylor Naval Ship R&D Center
MOORE, William T.	ARRADCOM
MOUNTER, L. A.	FSTC
NOBLE, Ben	U of Wisc (MRC)
NUNN, Walter R.	Center for Naval Analysis
PETERS, David A.	Washington U
PRESSMAN, Walter	CORADCOM
PU, San Li	Watervliet
RAJALA, David W.	U of VA
RALL, L. B.	U of Wisc (MRC)
REED, Harry L., Jr.	BRL
ROBINSON, Richard A.	Concepts Analysis Agency
ROSS, Edward W.	NARADCOM
ROSSER, J. Barkley	Univ of Wisc (MRC)
SCHLUSSEL, Kent	FSTC
SCHMITT, James A.	BRL
SCOTT, Brian R. (1LT)	BRL
SCOTT, Mary B.	FSTC
SHENOY, Pradash P.	U of Wisc (MRC)
SIMKUS, Anthony P. (COL)	ARO
SIMMONDS, James G.	U of VA
STEEVES, Earl C.	NARADCOM
SYMES, William W.	U of Wisc (MRC)
TAKAGI, Shunsuke	USA CREEL
TAYLER, M.	U of VA
THOMPSON, James L.	TARADCOM
VASILAKIS, John D.	Watervliet
WALTERS, David E.	ARRADCOM
WARD, Harold W.	U of VA
WARD, Jennifer	U of VA
WILSON, Stephen G.	U of VA
WHITE, Chelsea C., III	U of VA
WHITE, Michael W.	Army Sig War Lab
WHITEMAN, J. R.	Brunel U (UK)
WILLIS, Roger W.	TRASANA
WOLFF, Stephen	BRL
WONG, Eugene	U of Calif at Berkley
WU, Julian J.	Watervliet

NOTE: Approximately 50 additional people who did not register attended the talk given by Prof Kalman.

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER ARO Report Number 79-1	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) TRANSACTIONS OF THE TWENTY-FOURTH CONFERENCE OF ARMY MATHEMATICIANS		5. TYPE OF REPORT & PERIOD COVERED Interim Technical Report 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s)		8. CONTRACT OR GRANT NUMBER(s)
9. PERFORMING ORGANIZATION NAME AND ADDRESS		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS Army Mathematics Steering Committee on Behalf of the Chief of Research Development and Acquisition		12. REPORT DATE January 1979 13. NUMBER OF PAGES 496
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office) US Army Research Office DRXRO PO Box 12211 Research Triangle Park, NC 27709		15. SECURITY CLASS. (of this report) UNCLASSIFIED 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited. The findings in this report are not to be considered as official Department of the Army position; unless so designated by other authorized documents.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

This is a technical report resulting from the Twenty-fourth Conference of Army Mathematicians. It contains most of the papers on the agenda of this meeting. These treat various Army applied mathematical problems.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Stochastic models
Optimal predictors
Non-cooperative games
Shaped charges
Radial cracks
Stresses due to quenching
Symmetry and parity
Asymptotic solutions
Electromagnetic field penetration
Plastic deformation
Biharmonic functions
Large stiff systems
Integral Equations
Stochastic integral equations
Interpolating elastica
Linear Adaptive filters
Band matrices
Rotor blade dynamics
Bilinear stochastic systems
Hyperbolic conservation laws
A stochastic Reynolds equation
Fragmenting warheads
Efficient sets
Adaptive suboptimal designs
Software development in the Soviet Union
Martingale theory of random fields
Store and forward relays